DISSERTATION

Titel der Dissertation

„Regulation of Translation:

Small RNA and Messenger RNA – An Interplay“

Verfasser

Mag. Fabian Amman

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A 091 441
Dissertationsgebiet lt. Studienblatt:	Genetik - Mikrobiologie
Betreuerin / Betreuer:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Zu Beginn möchte ich mich bei den vielen Menschen, die mir diese
Arbeit zu schreiben ermöglicht haben, aufrichtig bedanken.

Zu allererst möchte ich Ivo und Xtof danken. Sie
haben mich in das weite und für mich zuerst neue
Gebiet der "computational biology" mit viel Geduld
eingeführt; hatten immer ein offenes Ohr für kleine
und große, für dumme und relevante Fragen.
Gleichzeitig haben sie mir den Raum und die Zeit
gelassen, meine Ideen zu verfolgen, und so meine
Erfolge und Misserfolge zu erarbeiten.

Auch haben sie mit dem TBI einen Rahmen geschaffen, in dem das Arbeiten ein
Vergnügen ist. Es ist diesem Klima geschuldet, dass eine kollegiales und
hilfsbereites Miteinander gelebt wird, was wesentlich zum Gelingen dieser
Arbeit beigetragen hat. Daher möchte ich Andrea, Athanasius et al.,
Christian, Florian, Heinz, Jörg, Judith, Mtw, Peter, Ronny, Stefan, Sven
und allen anderen für ihre Unterstützung und die gemeinsame Zeit danken.

Immerwährender Dank gilt meiner Familie, Mama, Papa, Simon,
Christoph, Verena und allen anderen; jeder von euch hat mich
auf seine Art, so lange er konnte, begleitet und unterstützt.
Ohne euch Allen und Allem, wäre diese Arbeit nicht.
Daher möchte ich sie euch widmen.

Danke.

Abstract

In the last decades the important role of small non coding RNA (sRNA) in bacterial post-transcriptional gene regulation was recognized. Meanwhile hundreds of sRNA genes are discovered, especially in the well studied model organisms, such as *Escherichia coli* or *Salmonella typhimurium*. Several experimental and theoretical techniques were developed to increase the accuracy of sRNA gene detection and to enable their annotation in a genome wide, high throughput manner.

In contrast to sRNA gene annotation, techniques for the characterization of the sRNA in the gene regulatory network are still in their infancy. In the course of this thesis, several existing gaps in the integration of computational methods into an efficient target prediction work-flow were identified and closed.

First, a statistical method to annotate transcription start sites from differential RNA-seq data is presented. In contrast to the boundaries of protein coding regions, hence the translation start and end position, the exact extent of the transcript itself is seldom known for bacterial genes. But since the most sRNA directly interact with the mRNA transcript, a detailed understanding of their architecture can be crucial.

Second, an efficient way to accurately calculate energetically favorable sRNA–mRNA binding sites is presented. Established algorithms to calculate the energy of putative sRNA–mRNA interactions either suffer from a low performance in reproducing known RNA interactions or are computational resource wise, very demanding. With `RNAplex`, part of the `ViennaRNA package`, it becomes possible to screen whole genomes for putative binding sites for a given sRNA.

And finally, a model framework is presented to calculate the effect of sRNA binding onto the translation initiation rate of a putative target mRNA. This model considers the physical interactions between the ribsome and the mRNA and the competing sRNA–mRNA binding. This way it becomes possible to test beforehand calculated interactions for their potential to inhibit or amplify translation, and shrink the number of predicted targets to a small set, which makes it possible to test them individually with more labor intensive in vitro or in vivo methods.

All three tools can be combined in a proposed sRNA characterization work-flow, and are helpful to make use of the limited experimental resources, to characterize sRNA and their mRNA targets, as efficient as possible.

Zusammenfassung

Die bedeutende Rolle die kleine, nicht Protein-kodierende RNA Moleküle (sRNA, vom englischen small RNA) in der Genregulation von Bakterien spielen, wurde erst im Laufe der letzten Jahrzehnte erkannt. Mittlerweile wurden Hunderte sogenannter sRNA Gene in bakeriellen Genomen entdeckt. Durch das kontinuierliche Entwickeln neuer Techniken ist die Zahl der bekannten sRNA Gene stetig am Steigen. Das Charakterisieren von bekannten sRNA wurde hingegen lange Zeit vernachlässigt.

Während der hier präsentierten Doktorarbeit wurden mehrere Lücken im systematischen, experimentellen Arbeitsablauf bisheriger Ansätze zur sRNA Charakterisierung identifiziert und durch rechnergestützte Methoden geschlossen.

Zum Ersten, für eine genau Analyse der sRNA–mRNA Interaktion müssen beide Komponenten möglichst detailiert bekannt sein. Für die mRNA mangelte es bisher meist an genauen Daten bezüglich des Transkriptionsbeginns bzw. -endes. Für diesen Zweck wurde eine statistische Methode entwickelt, um differentielle RNA Sequenzierungsdaten nach Transkriptions Start Positionen zu analysieren. Dadurch wird eine viel genauere Vorstellung über mögliche sRNA–mRNA Interaktionen gewonnen.

Zum Zweiten wird mit `RNAplex` eine Software, die Teil des `ViennaRNA package` ist, präsentiert. Der zugrunde liegende Algorithmus ermöglicht es mit hoher Genauigkeit, und dabei trotzdem sehr schnell, die energetisch günstigsten sRNA–mRNA Hybridisierungspositionen zu berechnen. Zuvor war es meist notwenig einen Kompromiss zwischen Schnelligkeit (und damit Anwendbarkeit auf ganze Genome) und Genauigkeit, einzugehen. Mit `RNAplex` gelingt es diesen Widerspruch zu lösen.

Obwohl gängige RNA Interaktionsprogramme sehr gut darin sind, bekannte Interaktionen präzise nachzuvollziehen, leiden sie meist unter einer geringen Spezifität, d.h. die Zahl der vorhergesagten Interaktionen ist bedeutend größer als die Zahl der Interaktionen, die sich experimentell als funktional erweisen. Daher wurde, zum Dritten, ein mathematisches Modell entwickelt, das es ermöglicht den Effekt einer vorhergesagten sRNA–mRNA Interaktion auf die Proteinproduktion von der gebundenen mRNA zu berechnen. Dadurch kann die Anzahl der Vorhersagen in einen Bereich gesenkt werden, der eine individuelle experimentelle Untersuchung jeder Interaktion mit in vitro oder in vivo Methoden ermöglicht.

Alle drei beschriebenen Methoden können sinnvoll in einem präsentierten Arbeitablauf kombiniert werden, mit dem es ermöglicht wird die Rolle von sRNA Genen in Bakterien zu charakterisieren. Dadurch können limitierte experimentelle Resourcen möglichst effizient für die Charakterisierung von sRNA und deren mRNA Partnern, eingesetzt werden.

Contents

I Preface 1

1 Motivation 3

2 Structure of this work 7

II Background 9

3 RNA 11
3.1 Introduction . . . 11
3.2 RNA structure . . . 13
3.2.1 RNA secondary structure . . . 14
3.2.2 Dynamics of secondary structure . . . 15
3.2.3 RNA tertiary structure . . . 17
3.3 Functional classification of RNA . . . 17
3.3.1 mRNA . . . 18
3.3.2 rRNA . . . 18
3.3.3 tRNA . . . 19
3.3.4 sRNA . . . 19
3.3.5 Ribozymes . . . 20
3.3.6 Other RNA species . . . 21

4 Messenger RNA 23
4.1 mRNA properties . . . 23
4.1.1 Transcription start site determination . . . 24
4.1.2 5' untranslated region . . . 27
4.1.3 Ribosome binding site . . . 28
4.1.4 Coding region . . . 30
4.1.5 Structural properties of 5' UTR and coding region . . . 31
4.1.6 3' untranslated region . . . 35
4.2 Gene regulation . . . 39
4.2.1 Transcriptional regulation . . . 39
4.2.2 Translation regulation . . . 42

5 Small RNA 49
5.1 Mode of action . . . 49
5.1.1 Regulation of protein activity . . . 50
5.1.2 Regulation of cis encoded mRNA activity . . . 50

5.1.3 Regulation of trans encoded mRNA activity 51
5.1.4 Bi-function sRNA 53
5.1.5 Hfq 53
5.2 Finding small RNA 54
5.2.1 Experimental discovery of sRNA 54
5.2.2 Computational discovery of ncRNA 55
5.3 Functional characterization of small RNA 57
5.3.1 Experimental discovery of sRNA targets 57
5.3.2 Computational discovery of sRNA targets 58
5.3.3 Confirmation of sRNA targets 58

6 Computational RNA Biology 61
6.1 In silico RNA folding 61
6.1.1 Secondary structure representations 62
6.2 RNA folding algorithm 64
6.2.1 Nussinov and Jacobson 64
6.2.2 Zuker and Stiegler 66
6.2.3 McCaskill 68
6.3 Local RNA folding 70
6.3.1 RNAplfold 70
6.4 Consensus structure for RNA alignments 71
6.4.1 RNAalifold 72
6.5 sRNA target prediction 74
6.5.1 Hybridization 74
6.5.2 Co-folding by concatenation of two sequences 75
6.5.3 Accessibility aware hybridization 75
6.5.4 Outlook 76

III Publication 79

7 Transcription start site annotation using dRNA-seq data 81
7.1 Statement of personal contribution 81
7.2 Article 81

8 Target prediction using conservation information 101
8.1 Statement of personal contribution 101
8.2 Article 101

9 sRNA binding and its influence on translation initiation 109
9.1 Statement of personal contribution 109
9.2 Article 109

IV Discussion 129

10 Discussion 131
10.1 Work-flow 132

10.2 Improvements & Outlook . . . 134
10.3 Future Application . . . 136
10.3.1 Medicine . . . 137
10.3.2 Biotechnology . . . 137

V Appendices 139

A List of Abbreviations 141

B List of Tables 143

C List of Figures 145

D Bibliography 147

E Curriculum Vitae 165

Part I

Preface

1 Motivation

One striking definition of **Life** is the concept of the seven pillars on which all living matter rests: Life needs to follow a kind of *Program*. For the life as we know it, this is implemented in the genomic DNA. The program must be flexible enough to allow for *Improvisation*, which means the program itself must be adaptable. A living organism must be confined into a limited space, separated from the surrounding environment. This *Compartmentalization* can be continued on smaller scale to separate different compartment within the organism. Life must be an open system metabolizing *Energy* to maintain itself and it must have the ability of *Regeneration*, to compensate the inevitable thermodynamic losses. The sixth pillar of life is its *Adaptability*. In an ever-changing environment improvisation is very often not enough to respond to sudden turns in the living conditions. Thus all living cells must have the possibility to run their program in different ways due to external stimuli. Finally, every living system must conduct chemical control and selectivity within its metabolism. This ability was termed *Seclusion* [119, 147].

In this work, I would like to present new tools to expand our knowledge of the sixth pillar: the Adaptability of bacteria gene expression. Living cells must adapt to sometimes drastic, sometimes only minor changes in their environment. Especially free living, unicellular organism, such as bacteria, which do not have the same possibility to maintain a body homoeostasis as multicellular animals, are exposed to an ever changing environment. The surrounding temperature is much less stable than the cells inner chemistry can tolerate to function optimal. The same is true for changes in salt concentrations, irradiation, toxic chemicals and the availability of nutrition, to name just a few. Most of these changes can be compensated by changing the internal program. Heat shock proteins can prevent other proteins to unfold or fold in an adverse manner due to a higher reaction temperature. If nutrients are very limited in the surrounding, the cell needs to built more transporters to exploit the resources as efficient as possible. At the same time it might be advantageous to use the available nutrients only for the most pivotal functions, reducing the energy needed for growth or reproduction to a minimum. On the other hand, when nutrient are plentiful it can be advisable to adapt the other way around: reduce the number of transporter, invest in growth, stock up reserves or mate.

The ability to adapt, presupposes the ability to sense the environment and the inner state of the cell and find a way to shut down or to boost, as a consequence, the appropriate function to the sensor stimulus. Mostly, enzymatic functions are executed by proteins, whose blueprint is encoded in the genomic DNA. Hence, shutting down or boosting is equivalent to deactivate or activate the protein production. Protein synthesis is a multi-step process which comprises the transcription of a genomic DNA section, i.e. a gene, into the so called messenger RNA (mRNA), which is bequeath to the translation machinery, i.e. the ribosome. Here a polypeptide chain is constructed, in which the sequence of amino-acids is determined by the sequence of nucleotide triplets. Each of the 64 possible nucleotide triplets corresponds to one of the 20 proteinogenic amino-acids or serve as so called stop codons. The translation product is a protein, which, in some cases, has to be activated. Finally the protein is degraded in a controlled fashion. All of the above mentioned steps, from the bare gene to the biological active protein is highly regulated, with the intent that the right amount of proteins, thus the right functional intensity, is present at the right time.

George Beadle and Edward Tatum proposed in 1941 their concept “that many biochemical reactions are in fact controlled in specific ways by specific genes” [13]. Since then, tremendous progress was achieved in finding mechanisms how gene products can control reactions and how genes themselves are controlled by other genes. A complex network of inter-dependencies emerged from the research of many scientists. This network consists of transcription regulation, translation regulation, mRNA and protein stability regulation and protein activation regulation.

In 1984, a new mechanism of translational regulation in bacteria was discovered by Takeshi Mizuno [153]. He could show that the translation rate of the bacterial outer membrane protein OmpF was sensitive to the presence of a small untranslated RNA. This RNA, later on it was named MicF, does not code for a protein but functions already as an RNA. It binds the mRNA ompF and leads to a decreased translation and eventually to the decay of its mRNA target. Since then, it could be shown that this mechanism, i.e. a small RNA hybridizes with an mRNA and modulates the translation rate, is wide spread in bacteria. Many pivotal cell functions seems to be controlled by small RNA. Most prominently, small RNA play a crucial role in the regulation of metabolism and virulence in many bacteria.

Meanwhile, in *Escherichia coli* more than 80 small RNA genes were successfully verified [180]. Computational screens based on sequence conservation, structural homology or expected components like promoters and terminators, suggest the existence of hundreds more [10]. The functional description of newly found sRNA genes becomes the main obstacle in broadening the existing gene regulation networks in bacteria. Functional characterization is still a challenging task. In contrast to miRNA in eukaryotes where a lot of binding rules are marked out [246], the interactions of sRNA with their mRNA counterparts show a striking variability in bacteria [10]. This is reflected by the fact that there is so far no satisfying stand alone technique to find new targets for sRNA. Experimental approaches are very labor intensive which means that they are not applicable to broad genomic screens (e.g. two-plasmid reporter gene assay [231]), or they are not suitable to properly distinguish between primary and secondary regulation effects (e.g. sRNA over-expression or deletion with downstream transcriptome profiling [199]).

Computational approaches focused so far on the thermodynamic hybridization properties of a given sRNA onto different potential mRNA targets. Although this proved to be very accurate in reproducing known interactions, this techniques suffer from a high false-positive rate, since any two sufficient ample RNA sequences will very likely show an energetically strong mutual binding site by chance.

Further on, to apply computational approaches which consider the mRNA structure a detailed understanding of the structure forming sequence can be essential. Until recently, only for a small set of interesting mRNA the precise transcription start site (TSS) was determined [66]. Recently high-throughput methods for TSS annotation were developed [200, 181], for which an automated, statistical sound analysis method is still lacking.

That is why, from my point of view, there is a great demand for new tools. On the one hand, to resolve the architecture of the transcription units and, on the other hand, to screen large target set for their potential to be regulated by a given sRNA. The later need to include other sources of information beside the common approaches of considering only the *one s*RNA – *one m*RNA system.

2 Structure of this work

The following work is structured in three main parts. Part II summarizes essential or interesting background information. First, a general introduction into RNA is given. The next chapter deals with mRNA in particular, presenting its property and how its role in the information flow from gene to protein is regulated on the level of transcription and translation. This leads to a chapter on sRNA. On the one hand, how post-transcription gene regulation by sRNA works. On the other hand, a summary of techniques to discover and characterize sRNA. The last chapter of the first part is dedicated to the basics of the vast tool box of computational RNA biology. Part II is meant to present the basis of and the scene around the story told and discussed in the following parts. Thereby, I dared to wander from the direct subject, especially if interesting details related to RNA and the importance of its structure are close-by along the road.

Part III presents three new contribution to the characterization of sRNA mediated post-transcriptional gene regulation. They are not ordered chronological according their publication date, but according the sequence how the developed tools can be applied in the course of the characterization of sRNA in a new species. In this sense, the first publication presents a tool to analyze data from the recently developed dRNA-seq technique. dRNA-seq aims to annotate in a high-throughput manner the transcription start sites in a bacterial genome. This is interesting for many application. For one, mRNA are the binding partner of sRNA, hence a detailed knowledge of the exact architecture of the mRNA and sRNA is of critical importance. Knowing the start site is already half the way to the finish line. So far dRNA-seq data were analyzed manually, introducing a lot of subjectivity with the cost of tedious and long error prone analysis. To change this, we developed a statistical approach to annotated transcription start site from dRNA-seq data in an automated manner. Details are given in chapter 7.

Chapter 8 presents an advancement of current RNA-RNA interaction prediction algorithms. `RNAplex` aims to merry the accuracy of detailed interaction calculation and the speed of less detailed interaction screening. Furthermore, this approach was extended to use the information of interaction conservation, increasing the specificity further.

Despite of the improvement of sRNA target prediction software over the past years, all of them, including `RNAplex`, suffer from a high false positive rate. They are quite successful in reproducing known interaction but report many non-functional interaction if used for de novo prediction. That is why we developed, and present in chapter 9, a model which aims to simulate the processes in the course of translation initiation with and without the involvement of sRNA. This model can be used to further evaluate predicted putative sRNA – mRNA interactions, for their potential to have an effect on translation initiation. Our model is semi quantitative in the sense that it also becomes possible to predict the type of regulation, whether positive or negative regulation, is caused by the particular sRNA mRNA interaction.

Finally, part IV summarizes the achievements accomplished in the course of this thesis and sets them into the bigger picture how a typical analysis of sRNA based gene regulation can look like and how the developed tools can be potentially refined and advanced in the future.

Part II

Background

3 RNA

3.1 Introduction

Ribonucleic acid (RNA) is, beside proteins and Deoxyribonucleic acid (DNA), one of the three major organic macro-molecules, which are essential for all known living matter. A few decades ago, its biological role was still disesteemed by reducing it to a simple intermediate in the flow of information from the storage (DNA) to the effector (protein). This concept was unintentionally consolidated in the *"Central Dogma of molecular biology"* [43, 46].

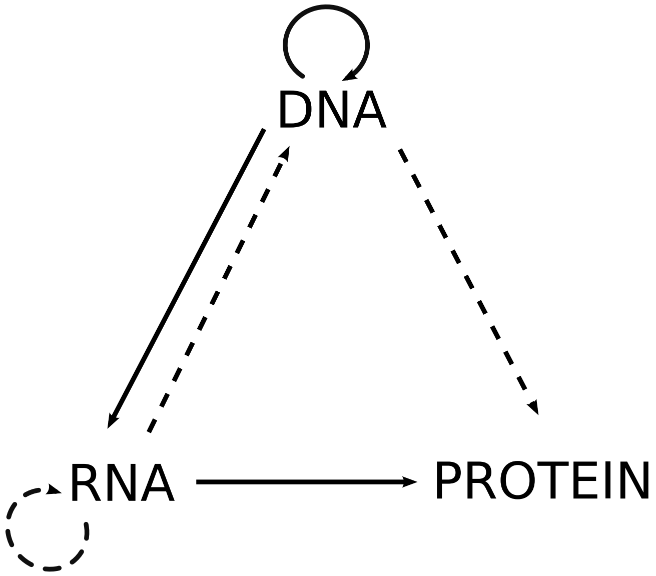


Figure 3.1: The "*Central Dogma of molecular biology*". Solid arrows indicate possible and probable transfers of information. Dashed arrows mark possible but unlikely transfers. Missing arrows were meant to be impossible transfers. This picture represents the knowledge of 1970 [46] which, in its pure interpretation, is still considered to be correct.

According to this credo, which was formulated by the highly respected and authoritative Nobel laureate Dr. Francis Crick, the DNA can reproduce itself and RNA can be produced from DNA. Proteins can be made from RNA templates. Other information flows, such as RNA → RNA, RNA → DNA, and DNA → protein were considered as theoretically possible but very rarely realized.

The central dogma is in fact valid till this day, but its common interpretations fall short to grasp the different roles RNA is known today to play. Thereby, the wide spread opinion which reduced RNA merely to a messenger of information from the DNA encoded genetic information to the important workhorse of the cell, the protein, was unintentionally cemented by the "Central Dogma of molecular biology" for decades to come.

Meanwhile, RNA and its function was revealed to be much more versatile. Beside the long known roles in transcription and translation, RNA was shown to be crucially involved in regulation and enzymatic catalysis of chemical reaction. In this respect, RNA seems to combine the ability to store information like DNA and to process information like proteins. The versatility arises from the combination of rather simple monomeric building blocks, comparable to DNA, and a complex structure, which can be formed with them. In this respect, RNA resembles proteins.

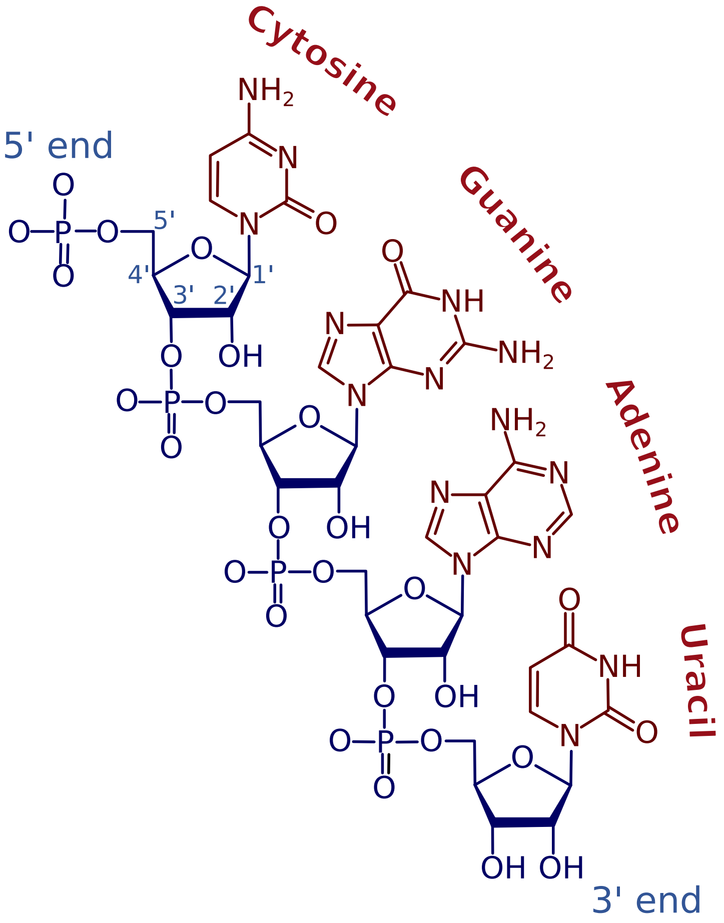


Figure 3.2: The chemical structure of an RNA chain. The sugar phosphate backbone is drawn blue. The attached bases, with the sequence CGAU in this case, are drawn in red. For the first sugar ring, the enumerated indices for the carbon atoms are labeled.

The basic building block of RNA is the nucleotide, which comes in four different "flavors". Nucleotide consists of a ribose sugar, one of four bases and a phosphate group. The ribose contains five carbon atoms, which makes it an aldopentose, whereas each carbon is enumerated from 1' to 5' (see Fig. 3.2). The base is attached to carbon 1'. In general, adenine (A), cytosine (C), guanine (G), or uracil (U) are utilized. Adenine and guanine are purines, containing two aromatic rings in a plane. Cytosine, and uracil are pyrimidines, with only one aromatic ring. The difference between cytosine and uracil, and adenine and guanine, respectively, are different functional groups attached to the basic framework of pyrimidines and purines. Different bases enable the RNA to form versatile structures, since A and U can interact to form a so called base pair via two hydrogen bonds, and G and C can pair each other via three hydrogen bonds (see Fig. 3.3).

DNA and RNA differ in two major aspects, with some important consequences. First, in RNA the base uracil is used. In contrast, in DNA the base thymine in incorporated. The second main difference is, as the name already implies, the usage of a different sugar. RNA has a hydroxyl group attached to the 2' carbon of the pentose ring, whereas DNA lacks this functional group[1]. This hydroxy group impairs the stability of RNA compared to DNA because it is more susceptible to hydrolysis.

Carbon atom 5' and 3' are connected to phosphate groups. This phosphate groups are able to bridge two nucleotides, forming the so called RNA backbone, a polymer with the uniform sequence *·phosphate·ribose·phosphate·ribose·phosphate·*. Since the bases are attached to the riboses, each RNA polymer can be described as an ordered, one dimensional chain of bases. This is called the RNA sequence, or sometimes the RNA primary structure. Since the riboses are connected via the 5' and the 3' C atom, the sequence is asymmetric, with many biological consequences. In general a sequence is given in a 5' to 3' orientation, which corresponds to the direction of RNA synthesis by the DNA dependent RNA polymerase. The commonly used terms "upstream" and "downstream" should be interpreted in this sense. Imagine a flow streaming in the direction of synthesis, upstream means in the direction of the 5' end, whereas downstream means into the direction of 3' sequence ends (see Fig. 3.2).

3.2 RNA structure

DNA normally appears double stranded in the cell, with two separate but complementary DNA molecules forming the iconic double helix structure with inter-molecular base pair. In contrast, RNA mostly occur single stranded, without a complementary partner strand[2]. This paves the way for intra-molecular base pairing within the same RNA strand. Single bases can pair with appropriate bases somewhere on the same strand, leading to a "tangled up thread". The pattern

---

[1] That's why it is called Deoxyribonucleic acid (DNA).

[2] A noteworthy exception are some viruses with a double stranded RNA genome. As a consequence, many species evolved a kind of molecular immune system, degrading dsRNA and corresponding sequences. This is applied in biotechnology, known as RNA interference [61].

of this interactions are called RNA structure. It was shown that many functions of RNA are critically dependent on its structure.

3.2.1 RNA secondary structure

RNA nucleotides can, due to their atomic structure, pair with each other via different interaction sites. The most common ones are the so called Watson-Crick base pairs. Thereby, hydrogen bonds are formed between uracil and adenine, and cytosine and guanine, respectively. Watson-Crick base pairs, also called canonical base pairs, are characterized by their isostericity, i.e., every pair has the same diameter, which allows to build regular helices. This distinguishes canonical from so called non-canonical base pairs.

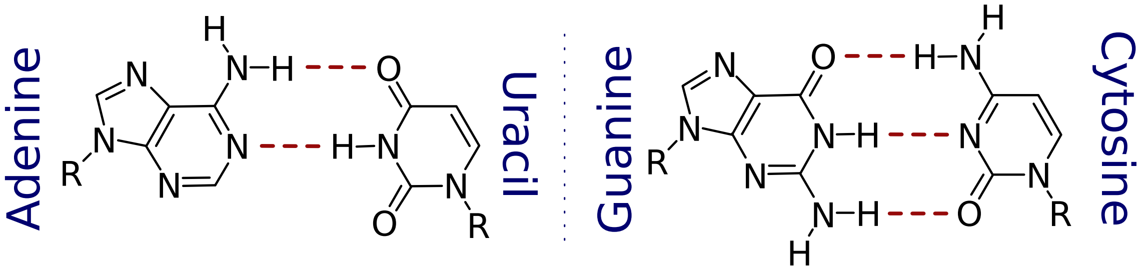


Figure 3.3: The anatomy of the RNA base pairs A-U (l.h.s.) and G-C (r.h.s.). The first is stabilized by two, the later by three hydrogen bonds between the charged groups.

In contrast to DNA, where the above mentioned base pairs are the only ones, in RNA molecules uracil and guanine can also pair with each other, forming a so called wobble pair[3]. In this case, the bases are linked via two hydrogen bounds. Wobble pair formation allows a higher diversity of structures for a given RNA sequence, and therefore has a great influence on RNA folding. For example, wobble pairs enable to reduce the number of different tRNA species in a cell. Instead of having one tRNA for each of the 61 possible amino acid encoding codons[4], in most systems one tRNA anti-codon can recognize more than just one codon, exploiting the pairing plurality of uracil and guanine [44].

Beside these canonical base pairs, other types of interaction between charged groups of the bases and of the attached ribose, contribute to the stability of the folded RNA molecule. To classify the variety of possible non-canonical interactions, the concept of different "edges" of the nucleotide was introduced [125]. All before mentioned canonical base pairs interact via the so called Watson-Crick edge with each other (see Fig. 3.3). From there, the shorter side in the direction to the hydroxy group at the 2' carbon of the ribose is called the sugar edge.

---

[3] Wobble pair differ from canonical base pairs with respect to the geometry of the pair. It is wider, hence interrupts the straight, rod shaped helix with a uniform diameter and introduces a wobble.

[4] Four different bases can be combined to $4^3 = 64$ different trinucleotide sequences. Three of which do not encode for amino acids but serve as stop codons.

On the opposite side is the Hoogsteen edge. Each of this edges contains functional groups which can interact with each other, forming non-canonical base pairs or even triplets [132, 133]. This further contributes to the variety of different structures which can be formed by RNA molecules.

An other important feature of RNA bases, which can contribute considerably to the stability of RNA structure, is base stacking. Due to the planarity of the bases itself and the planarity arrangement of two bases in a Watson-Crick base pair, two properly arranged bases, e.g. the bases in two consecutive base pairs, can interact with each other by electrostatic interaction, London dispersion attraction and shortrange repulsion [210]. How much stacking takes effect depends on the exact confirmation of the bases in their very local surrounding, considering neighboring bases and their relative positioning.

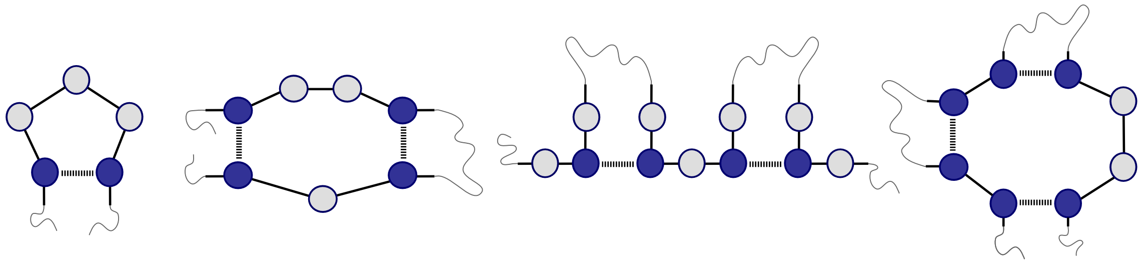


Figure 3.4: Loop types in RNA secondary structure. Solid lines represent the RNA backbone, dashed lines base paring interaction. Empty circles represent unpaired bases, filled circles represent paired bases. From left to right, the loops are names “hairpin loop”, “interior loop”, “exterior loop” and “multibranch loop”. Graphic adapted from [256].

The pattern of base pairs, i.e. which base interacts with which counter base, is called the secondary structure of the RNA. This can be classified in sets of different basic structure motifs. On one hand there are stems. A stem is a stretch of consecutive bases which form a double helix by Watson-Crick base pairing, similar to the well known DNA double helix. It consists of two anti-parallel, complementary sequence regions. Stems are separated by loops. According to the arrangement and the number of the adjacent stems they can be classified as hairpin loops, interior loops, exterior loops and multibranched loops (see Fig. 3.4).

3.2.2 Dynamics of secondary structure

Which secondary structure eventually will be adopted by a given RNA sequence follows the rules of thermodynamics. Accordingly, the most likely state of an RNA in the thermodynamic equilibrium is the one with the lowest free energy. Hence, it is often called *optimal structure* or *minimal free energy structure* (MFE structure). The free energy is a measure for the amount of work a system can perform, in other words the amount of total internal energy minus the amount of unusable energy. The later is expressed as the entropy of the system[5]. To calculate the minimal free energy of a given structure, one has to determine the total energy of the system and the entropy. Theoretically, this could be done for all possible structures, which would give the so called energy structure landscape of the RNA, where each point represents a distinct structure and is associated with its free energy.

Knowing the free energy E of a structure S directly provides the probability $P(S)$ that this structure is formed in the ensemble of all possible structures. This is expressed in the Boltzmann distribution

$$P(S) \propto e^{\frac{-E(S)}{RT}} \tag{3.1}$$

where, R is the gas constant and T the temperature. Using this equation, it can be illustrated that an RNA structure, even for a specific RNA sequence, is dynamic and temporally flexible. Therefore, let us consider an arbitrarily chosen structure from all possible structures for a given RNA molecule, one can easily imagine that adding or removing just a single base pair, leads to a tiny change in the energetic state of the system. According to Eq. 3.1, the likelihood of this altered structure is also only marginally changed. This consideration is valid for every possible structure, also the MFE structure. In other words, a given RNA sequence will fold and refold. The structure is dynamic, changing all the time. Thus, to describe the folding of an RNA by just one structure is a simplification. In some cases it is even an oversimplification.

Furthermore, a randomly structured RNA might have a very unfavorable energy, thus it will become more stable every time a random refolding leads to a lowering of the free energy. This process can be visualized as traveling on the afore mentioned energy landscape, whereas from each point the road downhill, the steeper the more, will be favored. Although, in the very long run this will lead to the absolute lowest point in the landscape, the way to get there can take a long time, much longer than the life time of the molecule itself, if the current location is separated by high barriers from the deepest valley. Therefore, one can assume that the functional important structure of an RNA can also be, due to the ephemerality of RNA in the cell, a complete different structure than the MFE conformation. This structure might have a low but not the lowest free energy.

In reality, the assumption of a random starting structure, is usually not true. In general, the RNA is always produced in 5' to 3' direction with the same speed for one transcript[6]. Already during transcription the newly produced 5' end can form base pairs. This can introduce an

---

[5] The energy that cannot be used to perform work is given by the entropy of a system multiplied by the temperature of the system.

[6] A notable exception are transcriptional riboswitches. In this case, the binding of an ligand to the riboswitch can alter the transcription speed of the polymerase, giving the nascent mRNA more or less time to form transcription termination hairpin. This is a well known example where not the MFE structure, but other semi-stable structure are functional important. Which of these structures are formed can be influenced by the ancillary conditions [83, 239].

important bias in the route of RNA folding, changing the set of adopted structures during the life span of the RNA molecule.

Already this short outline how RNA molecules fold into a functional structure indicates that the respective processes are far from being trivial. But in contrast to protein folding, which is computational still a very challenging task, the in silico prediction of RNA secondary structure is meanwhile well established and routinely done with accurate results. A more detailed introduction into computational RNA folding is given in chapter 6.

3.2.3 RNA tertiary structure

RNA secondary structure is the pattern of how bases interact to form helices connected by loops. These secondary structure motifs again can interact with each other, resulting in a three dimensional arrangement of the bases, which is called RNA 3D or tertiary structure. In contrast to RNA secondary structure, which is merely an auxiliary model to make RNA structure prediction feasible, the tertiary structure of an RNA represent the real physical structure of the molecule. The interaction on this level are mostly guided by non-canonical base pairs and soluble ions, e.g. Mg^{2+} can play an important role.

Kinetically it seems that the secondary structure is formed much faster, and only afterward tertiary structure is formed without much distortion of the secondary structure [28]. As a consequence, in silico RNA structure prediction is facilitated, since it can be assumed that it is not necessary to understand the whole 3D structure to predict the 2D structure. In contrast to the secondary structure, reliable tertiary structure prediction is less well established.

3.3 Functional classification of RNA

The understanding of the diverse functional roles played by RNA in the cell has dramatically changed in the last few decades. Not too long ago, RNA was seen as a mere vehicle to transmit the information stored in the DNA encoded genes to the protein executors. Then, messenger RNA and transfer RNA played the major role. Ribosomal RNA were also involved, but they were seen mostly as a scaffold to keep the ribosomal proteins in place to do the job. Meanwhile the picture has changed. RNA plays a crucial and much more active role in all this processes. Furthermore many new roles of RNA in the cell were discovered. In the following, some important classes of RNA coding genes will be briefly presented.

3.3.1 mRNA

An mRNA is the transcribed RNA equivalent of the DNA sequence between the transcription start and transcription termination site, and has distinct regions with different properties and functions. The pivotal part of an mRNA is the coding sequence (CDS), the section which eventually will be translated into the amino acid sequence of the encoded protein, which also distinguishes mRNA from other kind of RNA in the cell and are therefore called non-coding RNA (ncRNA). Unfortunately, the term non-coding is somehow misleading since it suggests that ncRNA do not code for a meaningful product. However, these genes only do not code for a meaningful amino acid sequence[7].

In contrast to eukaryotic mRNA, bacterial mRNA can possess more than one coding region per transcript. A coding region is marked by a start codon, a stretch of different size with no stop codon, i.e. open reading frame (ORF) and terminated by a stop codon. Upstream of the first CDS the 5' untranslated region (5' UTR) harbors many regulatory elements which are important for translation regulation. If on the same transcript downstream of the first CDS, one or more additional CDS follow, the mRNA is called polycistronic. Since in general the whole transcript is regulated as a unit, it can be advantageous to combine different genes into one mRNA. In the textbook example, this is applied, e.g., for proteins which are incorporated into the same protein complex, thus are always needed in the same stoichiometry.

As for the most RNA species, only recently the active role of the mRNA in transcription and translation was acknowledged. Chapter 4 deals in more depth with the processes and the importance of mRNA to establish and maintain homeostasis which is vitally important for all life.

3.3.2 rRNA

The largest portion of a total cell RNA extract consists of ribosomal RNA[8]. It can be found in all organisms making it one of the most fundamental constituents of life. The bacterial ribosome is composed of three different rRNA (16S, 23S, and 5S) and can be associated with more than 50 proteins [8]. It functions as the translational machinery, producing protein genes from DNA encoded blueprints. Furthermore it integrates many different signals to adjust the current protein production to the demand. For a long time, the rRNA was believed to function only as a scaffold to keep the ribosomal proteins in place. Meanwhile, the image has turned upside down: the central and active role of rRNA in translation was recognized, and for the proteins mainly scafold and regulatory functions remained. For example, the rRNA actively

---

[7] That is why some authors proposed the term "functional RNA" (fRNA) instead, e.g. [9, 161], which somehow suffers from the similar vagueness since an mRNA is also functional.

[8] In Yeast 80 % of the total RNA are different rRNA [241].

catalyzes the chemical attachment of additional amino acids to the nascent polypeptide (see page 20). Furthermore, the ribosome recognizes the mRNA and its coding region, making the ribosomal RNA and its interactions a key party in post transcriptional gene regulation (see section 4.2.2 and chapter 9).

3.3.3 tRNA

The second most frequent type of RNA in the cell is the transfer RNA. The tRNA is typically between 73 and 94 nt long and works as an adapter between the mRNA and the protein by translating the nucleotide triplets of the mRNA into the amino acid sequence of the protein in an unambiguous way.

The secondary structure of a tRNA can be drawn in the typical cloverleaf form (Fig 3.5). On top the 5' and 3' ends are connected via a stem. The amino acid is bound to this stem, therefore its name, acceptor stem. Furthermore, there are the D-loop, the anti-codon loop, the variable loop, and finally the TΨC-loop (listed from the 5' to the 3' end). Each of the loops has a specialized function. The anti-codon loop harbors the anti-codon. Here, the tRNA interacts with the mRNA's codons, giving the tRNA its specificity in the genetic code. It is also important to stabilize the whole mRNA·tRNA·rRNA complex in the course of translation initiation [155] (see also leaderless mRNA and translation initiation, on page 29 and 43, respectively). The D-loop acts as a recognition site for the right aminoacyl-tRNA synthetase to the correct tRNA, mediating the specificity to charge the tRNA with the correct aminoacid residue [86]. The T-loop is involved in the recognition of the tRNA by the ribosome [233].

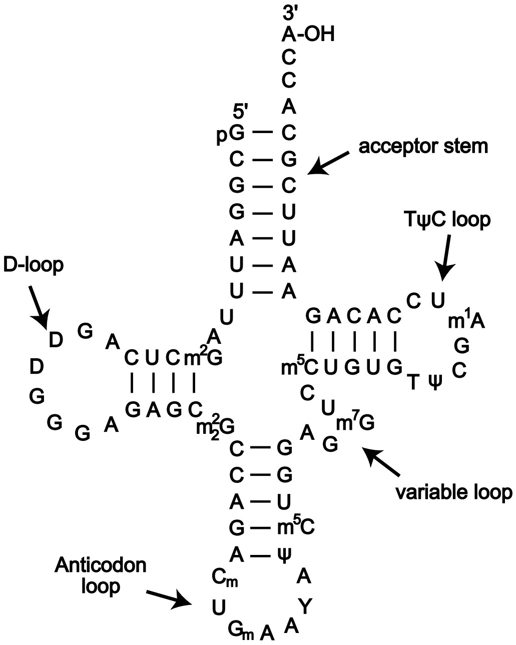


Figure 3.5: An idealized secondary structure of a typical tRNA. Source: Wikipedia.org User:Yikrazuul

3.3.4 sRNA

The term sRNA is often used somehow ambiguously. In former times it was used for tRNA, as an abbreviation for "soluble RNA". Meanwhile, it often means "small RNA" in the literature, or sometimes sRNA refers to different RNA species, often in the general sense of non-coding RNA, e.g. in [248]. Throughout this work, the name sRNA is used for *bacterial small trans-encoded anti-sense RNA*. This bulky term summarizes many characteristics of sRNA.

- This kind of RNA mediated gene regulation mechanism is unique to **bacteria**. Eukaryotic systems have in spirit similar strategies, but the details differ substantially.
- **Small** refers simply to their size (up to 200 nt for a typical sRNA [203]) which has to be seen in relation to mRNA, which are typically longer.
- **Trans-encoded** mean that the sRNA gene is transcribed into the functional small RNA, which diffuses through the cytoplasm, potentially interacting with all possible mRNA and proteins. This distinguishes sRNA from other RNA based translation regulation mechanisms, such as riboswitches, since they are physically attached to their regulated gene.
- The term **anti-sense** describes already the main mode of action. The sRNA can bind its target interaction site by base pairing with a small stretch of complementary sequence. In similar contexts, anti-sense can also refer to the encoding position, meaning the sRNA gene itself is encoded on the opposite, hence anti-sense, strand of the target gene (see section 5.1.2). Here, this denotation is explicitly not meant.

sRNA play an important role in post-transcriptional gene regulation. So they do in this thesis. Therefore, chapter 5 gives more detailed background information on the role and mechanisms important for sRNA regulation. Chapters 8 and 9 give the author's contributions in the field of sRNA target prediction.

3.3.5 Ribozymes

Maybe the most surprising insight into the functional repertoire of RNA was the discovery that RNA itself, without any protein contribution, can act as an enzyme with catalytic properties. This led to a radical change of how RNA was seen, and was awarded with the Nobel price in Chemistry in 1989 to Thomas Cech and Sidney Altman. This discovery did not just push the protein from the pedestal of being the unique organic compound with enzymatic activity, it also served as food for exciting thoughts on the origin of life and whether this duality of encoding DNA and enzymatic protein might be a successful specialization of a world where both functions were exclusively united into one chemical family [70].

From a historical point of view, three distinct examples of ribozymes deserve special mentioning here: RNase P was, beside self-splicing of group I introns [121], the first discovery of an RNA with enzymatic activity [215]. It is conserved from bacteria to human and consists of a protein·RNA complex. Its enzymatic activity was shown to be executed by the RNA alone, although with reduced efficiency. The best conserved function of RNase P is the maturation of pre-tRNA by processing the 5' leader sequence. In contrast to most other ribozymes, RNase P does not recognize its substrate via complementary base pairing but by a mechanism using a metal ion, hence in this aspect the mechanism resembles more proteins enzymes [157].

The probably most used and thus most essential ribozyme in nature is the peptidyl transferase 23S rRNA. In all organisms, the pivotal part of translation is the peptide bond formation between peptidyl-tRNA and aminoacyl-tRNA. This reaction is carried out in the peptidyl transferase center of the ribosome [188]. Although the ribosome is a large complex of many RNA and proteins, the enzymatic activity is accomplish by the RNA alone, making the protein portion responsible for structural arrangements and regulation [35].

Last but not least I would like to mention the ribozyme glmS. The glmS mRNA codes for an enzyme which converts fructose-6P into glucosamine-6P. In the untranslated region upstream of the coding region start, several different RNA regulatory sites can be found, making it a showcase example of RNA mediated translation regulation. On the one hand, in gram negative bacteria such as *Escherichia coli*, glmS is regulated by small RNA. GlmZ activates glmS translation upon binding [230]. On the other hand, the glmS untranslated region of gram positive bacteria harbors a riboswitch which selectively binds glucosamine-6P. Subsequent to ligand binding, the RNA cleaves itself which leads eventually to the inactivation and degradation of the glmS transcript [145].

3.3.6 Other RNA species

Once the important role RNA can play was acknowledged, research expanded the understanding of RNA-involved processes in all directions. As of 2012, the non-coding RNA family database (Rfam) listed 2208 different RNA families with more than 6 million sequence entries [31]. Although, the majority are unique to eukaryotes, many are found in bacteria (see tab. 3.1).

Table 3.1: Number of Rfam families per taxonomic domain. Since some of the total 2208 Rfam families are not unique to one domain, the numbers do not add up to the total 2208 described families [31].

Domain	Total Number	Unique Number
Archaea	88	76
Bacteria	462	427
Eukaryotes	1323	1294
Viruses	161	138

Exemplarily, two shall be presented in more detail. A recent discovery of an RNA mediated system is the so called CRISPR system (Clustered Regularly Interspaced Short Palindromic

Repeats). Bacteria cells are under constant pressure from bacteriophages[9]. Therefore, around 40 % of sequenced bacteria were shown to posses CRISPR loci, an immune system like defense mechanism against phages, in which RNA plays an important role [94]. Intruding phage DNA is processed, producing small fragments which can be incorporated into the CRISPR locus in the chromosome. These alien sequences are constitutively expressed producing small RNA molecules, which guide exonucleases to degrade foreign nucleic acid, DNA and RNA alike. The exact mechanism how this is achieved and regulated still remains to be elucidated [94]. However, it is noteworthy that for a long time the bacterial small RNA system was seen as an analog to the eukaryotic microRNA system. Meanwhile it became clear that the CRISPR system represents a much more mechanistically related (though not necessarily evolutionary related) mode of action. Nevertheless, this related system is used for a quite different purpose, namely foreign nucleic acid defense instead of post transcriptional gene regulation [136].

One of the more odds are tmRNA. Transfer-messenger RNA are a hybrid with a portion resembling a tRNA and the other part is messenger RNA like. In bacteria, tmRNA are associated with one protein, forming a ribonucleoprotein complex [182]. This complex releases ribosomes from mRNA, if the conventional mechanism of translation termination failed, e.g. due to the lack of a stop codon. Normally this would lead to a stable, nonfunctional mRNA·ribosome complex. In order to reactivate this stuck ribosomes the tmRNA binds the ribosome in the same way a regular tRNA would do, providing an mRNA-like segment with an short open reading frame (ORF). This ORF is then translated, in a process called trans-translation, until the ribosome is released from the auxiliary coding region. The dysfunctional protein contains a tag, encoded by the tmRNA ORF which marks the protein for degradation. tmRNA are a beautiful example how RNA can mimic other structures to adopt their functionality and how different RNA can be interconnected in a module like manner [182].

---

[9] It was estimated that there are more than 10^{30} phages in the ocean [137], outnumbering bacterial cells by a 1 to 10 ratio [41].

4 Messenger RNA

Messenger RNA play a pivotal role in the process of gene expression. They link the information flow from the gene to the protein. Formerly, is was considered a passive container of information. Meanwhile, the active role which RNA fulfills in many regulatory processes is acknowledged. In eukaryotes the most prominent and well understood example is splicing. Pre-mRNA undergo different maturation steps and rearrangements to produce the blueprints for the required proteins. This opens the opportunity to use the same gene although the gene product needs specialization in different tissues or different developmental stages [117]. For bacteria splicing of some genes was reported but generally bacterial mRNA is not spliced [59]. Instead, other mechanism evolved to optimize the gene regulation, taking action both on the level of transcription and translation. For this, structural and sequence properties of the different parts of the mRNA are important. Understanding an mRNA with all its features enables us to see its role in the orchestrated protein synthesis, providing the right amount of the right effector to the right time.

4.1 mRNA properties

Bacterial mRNA are the RNA copy of a part of the genomic DNA. For this the DNA coding strand is duplicated by using the DNA template strand as a blueprint. Transcription starts at the so called transcription start site (TSS) and runs until the transcription terminator is reached. In between, there are one or more so called coding sequences (CDS). As a result each bacterial mRNA can possess following features (also see Fig. 4.1).

5' untranslated region are between the transcription start and the translation start of the first CDS. 5' UTR can also be missing when the translation start site is equal to the transcription start site, which is not uncommon.

Coding sequence are the open reading frame whose nucleotide triplets code for the amino acid sequence of the protein. Bacteria mRNA can be monocistronic (like eukaryotic mRNA) with just one CDS per mRNA molecule, or they are polycistronic, harboring more than one CDS. The different CDS can be separated by an untranslated region or even overlap with each other.

3' untranslated region form the end of the mRNA, between the last translation stop and the transcription stop.

Each of these regions play a distinct role in the process of regulated translation. Therefore, they can harbor special structures or sequence motifs. In the following section I'd like to shed light on the nature of the different mRNA part and the techniques to study these.

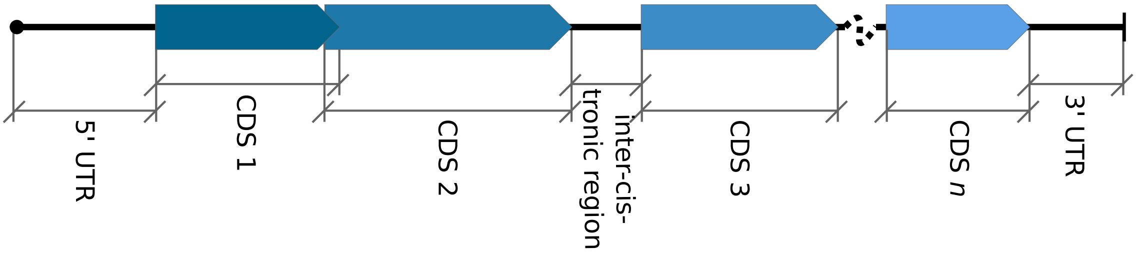


Figure 4.1: Possible arrangement of a polycistronic mRNA. The 5' untranslated region is followed by the first coding region. Coding regions can overlap, directly follow on each other without a gap or be separated by an untranslated segment (intercistronic region). The last stretch of sequence behind the last coding region is called 3' untranslated region.

4.1.1 Transcription start site determination

A comprehensive knowledge of the TSS positions in a genome is an important, although up to now, neglected part of thorough bacteria genome annotation. This is because many processes can only be understood if TSS are known. Some because they directly affect TSS positions, e.g. the promoter organization, or because they directly depend on the TSS positioning, e.g. the sequence and hence the structure of the 5' untranslated region.

For eukaryotic models, such as human or mouse, experimental high-throughput methods to detect TSS were developed already ten year ago [204]. Due to the different organization of prokaryotic transcripts, mainly the lack of a 5' cap structure, these techniques can not be applied to bacteria. Here, different methods were introduced, which will be briefly reviewed in the following section.

Transcription factor binding site annotation Promoter position are tightly constrained relative to the transcription start sites. In *E. coli* the distance is typically 7 to 10 nt [149]. Experimentally, promoter can be detected by evaluating DNA segments with a binding affinity to a transcription factor with DNA binding ability. The most widely applied technique is the so called chromatin immunoprecipitation (ChIP) combined with some sort of read out procedure. For this, genomic DNA and all its bound proteins are cross-linked. Subsequently, the DNA is fragmented by sonication. The crucial step follows by pulling down the protein of interest with its attached DNA. This is done with specific antibodies. In former times, this way enriched DNA, which is associated with the DNA binding TF, was analyzed with qPCR. Since this becomes impossible for a genome wide analysis, later on the DNA read out was performed by micro-array chips (ChIP-chip) or DNA sequencing (ChIP-seq) [30].

To circumvent the need of specific antibodies against the TF of interest, which are oftentimes difficult to produce, ChIP-chip can also be modified. For example [85] showed that recombinant TF with an attached tag, which allows the usage of affinity media instead of antibodies. Thereby, the experimental procedure is simplified and thus its possible application broadened. This technique is called ChAP-chip (chromatin affinity purification).

Computational promoter prediction Since promoters have some more or less well defined features, which distinguish them from the background DNA sequence, several approaches attempt to annotated their position in-silico. For that purpose, information from characteristic sequence motifs [49, 72] and thermodynamic properties were successfully exploited [107, 183, 221].

The later uses several details shared between the DNA within promoters, i.e. low stability, high curvature and less bendability [108]. All of these features seem to ease the binding of the transcription factor and the opening of the transcription initiation bubble. Thereby, the low stability which is equivalent to a low melting temperature (the energy needed to separate the two DNA strands) is the easiest to deduced from the DNA sequence.

Although the described approaches to annotate promoter and TSS from in silico analysis work well in reproducing known TSS, their use as de novo annotation tool is limited by a high false positive rate [60, 107, 171]. Thus, reliable understanding of transcription architecture and transcription regulation still requires sound experimental techniques.

TSS annotation Beside the afore mentioned methods which use characteristics of the genomic DNA to describe potential promoters and their corresponding TSS, the annotation can also be approached from the transcript side. In a first step the physical boundary of the mRNA is determined, either in terms of nucleotide sequence or simply in terms of distance from a well annotated feature, such as the translation start. After wards this can be tracked back to the genomic DNA, providing the position of transcription initiation in genomic coordinates. To identify the exact 5’ end of the transcript, different approaches were successfully applied, primer extension and RACE, being well established and very accurate techniques and, very recently, dRNA-seq.

Primer extension One of the first techniques to locate the 5’ end of RNA was so called Primer extension. First, radioactive labeled primers against a well described position at the 3’ end are used in a PCR like reaction. Since the reverse primer is not known and thus not used, no amplification takes place. Still, after separating the RNA in an acrylamide gel, due to the sensitivity in detecting the radioactive labeled RNA, the length of the mRNA from the 5' end to the primer hybridization position can be deduced [222].

RACE A wide spread technique to find the exact boundaries of a transcript whose sequence is only partly known is the so called rapid amplification of cDNA ends (RACE). Generally, RACE can be used with minor adaptations to annotate the 5' and 3' ends. To detect the 5' end a PCR primer complementary to the coding sequence of the gene of interest (GOI) mRNA is designed. This primer points to the 5' end, resulting in a cDNA fragment from the primer binding site to the original TSS, where due to lack of template the polymerization stops. To amplify this fragment in further PCR cycles, a poly A tail is appended using terminal deoxynucleotidyltransferase (TdT) and dATP. The poly A tail and a complementary second primer is used to specifically amplify the product and thus the sequence of the 5' UTR [65]. The read out is classically done by sequencing of the amplified segments [165].

The application of a refined version of RACE was described in [149]. There, the *E. coli* genome was exhaustively tested for TSS, revealing that the transcriptional architecture is indeed much more complex than the simplified textbook image implies. Although the technique showed its usefulness, it still cannot distinguish between primary TSS originating from transcription initiation or secondary TSS created by RNA processing.

dRNA-seq To overcome this problems, similar to the cap in eukaryotic cells, a distinct mark, separating primary and secondary 5' mRNA ends are needed. This was found in the characteristic phosphorylation pattern of primary transcription starts. The mono-nucleotides for transcription are provided to the polymerase in the form of nucleotide triphosphates. In the process of transcription elongation the triphosphates are broken down and the released energy is used to form a phosphodiester bond between the newly conjoined nucleosides. As a consequence, the first nucleotide still has its three phosphates attached at the 5' carbon atom. In contrast, if the phosphodiester bond of two consecutive nucleosides is broken due to endonucleolytic cleavage, the remaining fragment is a 5'-phosphomonoester.

This difference between primary transcription start sites and secondarily introduced transcript starts in combination with deep RNA sequencing can be deployed to design experiments to annotate genome wide the 5' ends of all currently expressed transcripts. For this purpose, two approaches were independently developed.

Sorek et al. [250] introduced a method which uses the enzyme *Tobacco Acid Pyrophosphatase* (TAP) to remove two phosphates from the 5' triphosphate nucleotides. This way, the resulting fragments are a possible substrate for the following sequencing adapter ligation, which is applied. Eventually, a strand specific sequencing library is constructed. Relative to a library with no TAP treatment, the reads whose 5' end corresponds with an authentic transcript start, are overrepresented in the TAP treated library.

A similar approach was developed by Sharma et al. [200]. Here, instead of TAP the enzyme *Terminator-5'-phosphate-dependent exonuclease* (TEX) is used. TEX specifically degrades RNA with a 5'-monophosphate. Transcripts with a 5'-cap, 5'-hydroxyl group or, most interestingly in our case, a protective 5'-triphosphate are spared. Similarly to the TAP based approach, this leads to an enrichment[1] of reads associated with primary transcription starts in the TEX treated library compared to an untreated library.

The read-out of both methods is similar. The libraries are sequenced, the produced reads are mapped back to the reference genome, and finally, positions with an enrichment of read starts in the treated library compared to the untreated library are identified as transcription start sites. This last step bestows the name on this method. Since the read-out is based on the relative difference between two libraries, the technique was named differential RNA-seq (dRNA-seq).

dRNA-seq was successfully applied to several different organism, e.g. [194, 149, 123]. Many single predicted TSS were confirmed by individual approaches such as RACE or primer extension, showing the method's accuracy. The main difficulty to apply this method is still the analysis of the huge amounts of produced data. So far most studies simply visualized the libraries in a genome browser and eyeballed every position for a putative signal. This is not just very labor intensive but also rather subjective, introducing ambiguity into the analysis. Because of this, a new automated method of dRNA-seq data analysis is proposed in chapter 7.

4.1.2 5' untranslated region

Once the exact position of the transcription start site is known, the precise extent of the 5' untranslated region can be deduced.

Previous analysis in eubacteria and archaea bacteria showed a somehow nonuniform picture of 5' UTR length distribution. For *H. pylori* around 4 % of the annotated TSS correspond to leaderless mRNA[2]. Only very few have a length between 10 and 20 nt. The majority of genes show a transcription to translation start distance of around 30 nt. After that, the distribution slowly flattens, whereas single UTR with more than 400 nt could be observed [200].

In *Xanthomonas campestris pv. vesicatoria* the results are similar, but there are even more leaderless mRNA (∼14 %) and the majority of 5' UTR had a length of 20 to 25 nt [194]. For *Escherichia coli*, most UTR have also a length of 20 to 30 nt but there leaderless mRNA appear

---

[1] In fact, it leads to a depletion of reads which are not associated with a primary transcription start.

[2] Here, leaderless mRNA were defined as mRNA with a 5' UTR with less than 10 nt length. A total of 34 out of 825 TSS felt into this category

to be much less common [149]. *Listeria monocytogenes* and *Listeria innocua* have also hardly any leaderless mRNA and a median UTR length of 33 nt [251]. *Thermotoga maritima* shows a bimodal UTR length distribution, thereby most genes have a preceding UTR of either 11 to 17 nt or of 26 to 32 nt. Only a small fraction of UTR do not fall into this ranges [123]. The situation completely differs for the archaea *Sulfolobus solfataricus P2*. Here, the vast majority of all genes are expressed leaderless. A distinct 5' UTR seems to be the exception and not the norm [250].

The significance of the 5' UTR and also of knowing the exact coordinates of this region lies in its important role in gene regulation. Many special structures such as riboswitches, but also binding sites for sRNA and the ribosome itself, are situated in the 5' UTR. Hence, knowing the sequence of the untranslated region facilitates to interpret possible effects of putative functional entities, i.e. structures or sequence motifs. This is a prerequisite to fully understand gene regulation and gene expression.

4.1.3 Ribosome binding site

To fulfill its purpose, the mRNA must be enabled to pass on its decoded information to be used to synthesize a protein. This is executed by the ribosome (see section 4.2.2), whose interaction with the mRNA is stabilized mainly via two sequence motifs, the Shine-Dalgarno sequence and the start codon. Therefore, the ribosome – mRNA hybridization is very much influenced by the accessibility and the sequence of these two regions. The ribosomal structure is well defined, the mRNA in contrast, shows great structural diversity. Section 4.1.5 deals with this aspect in more details.

Shine-Dalgarno sequence The Shine-Dalgarno sequence, or for short SD sequence, is an around eight nucleotide long sequence stretch upstream of the translation start. The ribosome includes on its 3' end a so called anti SD sequence. The complementarity between the SD and anti SD sequence strongly influences the strength of ribosome binding and eventually the efficiency of translation. As illustrated in table 4.1, in *Escherichia coli*, for example, it could be shown that the SD sequence UAAGGAGG is roughly four times more efficiently translated than the same gene with the truncated SD version AAGGA [185].

Later results showed that there is an optimal SD length. The picture of, the longer the SD the higher the translation rate, does not hold a critical examination. This might be explained by the fact that the SD – anti SD interaction must be resolved later on in the course of shifting from initiation phase to translation elongation phase. A too stable ribosome – mRNA interaction might slow down this process and blocks the ribosome binding site (RBS) for other ribosomes [167].

Table 4.1: Different SD sequences and their alignment with the anti SD sequence from the 16S ribosome. The SD with the more extended complementarity shows a four-fold increased translation rate [185].

anti SD	3'	`--AUUCCUCC--`	5'
long SD	5'	`--UAAGGAGG--`	3'
short SD	5'	`---AAGGAaa--`	3'

Start codon Beside the SD sequence the ribosome complex specifically interacts with the mRNA via the anti codon on the tRNA in the ribosomal P-site and the start codon. For *E. coli*, the UCSC genome browser lists 2391 annotated start codons [151]. 2160 (90.3 %) of them correspond to the nucleotide triplet AUG. With 159 (6.6 %)and 41 (1.7 %) examples, the triplets GUG and UUG, respectively, are much less common. The remaining 31 are other triplets, none of which is observed more than twice [21].

Since translation initiation always starts with the same tRNA fMet-tRNA$_f^{Met}$, with the anti codon 3'–UAC–5', the codon – anti codon interaction, hence the complete translation initiation complex, is the most stable one, if the 5'–AUG–3' start codon is provided. This is also reflected in the increased translation efficiency for genes starting with this particular codon [185].

Spacing The ribosome contacts the mRNA at its SD sequence and the start codon simultaneously. Since the relative position of the anti SD sequence and the anti codon of the fMet-tRNA$_f^{Met}$ in the ribosomal P-site is inherently given from the well conserved ribosome 3D structure, the spacing between the more diverse positions of the SD sequence and the start codon also influences the translation efficiency. Experiments in *E. coli* showed that the peak in translation efficiency corresponds with a distance of 5 nt, dropping to a relative translation rate of ~50 % when the distance is increased to 9 nt [37].

S1-binding site Some genes are translated even in the complete absence of a functional SD sequence. For one, leaderless mRNA has to be mentioned here. But there are also genes with a distinct 5' UTR still lacking the SD which are successfully described in *E. coli*, e.g. some plant viral mRNA such as alfalfa mosaic virus RNA 4 and tobacco mosaic virus RNA [227].

Therefore, it was proposed and tested that a A/U-rich sequences in front of the SD sequence works as a translational enhancer. The ribosomal protein S1 specifically binds the mRNA in a site specific manner, whereas there is no strict sequence motif described yet [24, 227].

Leaderless mRNA Beside the afore described general architecture of canonical mRNA, there are notable exceptions to this scheme. The most prominent ones are the so called leaderless mRNA. This mRNA species lacks the 5' UTR and starts directly with the start codon AUG at its 5' end. Since for canonical mRNA the ribosome binding is guided by motifs positioned upstream of the start codon, at first the exact mechanism how leaderless mRNA are efficiently translated remained unclear (e.g. [104]). In this phase the idea if the so called downstream box was resurrected (see [202] and page 30), but was later on dismissed [155]. Meanwhile, several mechanism were proposed how to efficiently initiate translation from leaderless mRNA. Similar to canonical mRNA the structure, or the lack of structure, in the translation initiation region, in this case the beginning of the coding sequence, seems to be important [139]. The established picture of leaderless mRNA translation initiation assumes that instead of the pre-initiation complex formation, the pre-assembled 30S ribosome including the bound initiation factor IF2 and fMet-tRNA$_f^{Met}$ binds the mRNA, whereas no other signal than the canonical AUG start codon is needed [155].

One notable function of leaderless mRNA are associated with the response to stress. This is mediated by a sophisticated mechanism, involving the endonuclease MazF. There, MazF selectively cleaves mRNA close to the start codon, rendering them into leaderless mRNA. Additionally, MazF removes 43 nt from the 3' end of the 16S rRNA. This part includes the anti Shine-Dalgarno sequence, which is essentially for rRNA binding onto canonical mRNA. This process creates a ribosome sub-population which is impaired in transcribing canonical mRNA and thus transcribes preferentially leaderless mRNA. At the same time leaderless mRNA are produced by the same mechanism in a selective manner [235].

4.1.4 Coding region

The coding region is the part of the mRNA which is translated into the protein. The before mentioned start codon marks the start and is already part of the coding region. Beside this landmark codon there are a few other, important sequence feature.

Down stream box Beside the classical factors important for efficient translation initiation, other, less important motifs were proposed. One, which lead to controversial discussion over a long period of time, is the so called downstream box (DB) [211]. The DB is assumed to be positioned 3' from the initiation codon, whereas the exact position is gene dependent, i.e. less conserved than for example the SD-sequence position. It shows some complementarity to the rRNA at position 1469 to 1483. Consequently, this region was named anti-downstream box (aDB) [143]. Its functional role could be shown by deletion and mutation analysis [212]. On the other site, reasonable doubt arose about the importance of the DB sequence and the validity of the proposed aDB – DB mechanism. The skepticism is mainly based on contradictory results from structural analysis of the ribosome complex, indicating that a simultaneous binding of the aDB – DB and the initiator-tRNA to the start codon is sterical not feasible [154].

Stop codon Translation termination is triggered by different nucleotide triplets on the mRNA sequence. In contrast to coding triplets (codons), this stop triplets are not recognized by a tRNA. Instead, prokaryotes possess so called release factors (RF) which interact with the mRNA, the ribosome and the preceding tRNA in the P-site with the nascent polypeptide chain [196, 255]. This causes hydrolysis of the ester bond of the peptidyl-tRNA and the release of the ready-made protein from the ribosomal complex. Each release factor recognizes different stop codons. RF-1 interacts with UAG. RF-2 interacts with UGA. The triplet UAA is accepted by both release factors [45].

4.1.5 Structural properties of 5' UTR and coding region

The connection between the structure of the RBS and the translation efficiency has been discovered long time ago [100, 247],and repeatedly confirmed (e.g. [82, 109, 50]). The well-established ratio behind is the need of the ribosome to compete with possible internal base pairs to get access to the mRNA and assemble into the translation initiation complex. In agreement with this consideration, in bacteria, the region of 30 nt in front of the translation start is significantly more accessible compared to randomly shuffled sequences (Fig. 4.2 top lane). It must be emphasized that the accessibility differences are only observable if the sequences are nucleotide shuffled. This destroys the dinucleotide content. If the sequences are shuffled in a way to conserve the dinucleotide composition [101], for the UTR the differences between the native sequence and the shuffled sequence are negligible.

A more detailed inspection of exemplary bacterial species (Fig. 4.4.A-G) shows a typical pattern of base accessibility. A few dozen bases in front of the start codon the accessibility starts to increase steadily, to reach a maximum between -20 to -10 nt relative to the translation start. The bases directly around the start codon show a relative drop in accessibility, followed by a new maximum downstream of the CDS start. The point of discontinuity at the translation start might be caused by the predominant start codon AUG itself. Interestingly, this first peak in accessibility does not fully correspond with the location of the SD sequence, which is normally seen as the anchor for the whole RBS.

The increased accessibility of the mRNA bases around the translation start position has major implication for sRNA regulation in general and sRNA target prediction in particular (see section 6.5). On one hand, the fact that mRNA 5' UTR are more accessible eases the binding for the ribosome, is also true for the binding of sRNA. Moreover, sRNA mostly interfere with translation initiation, and the RBS is certainly the place to do so. Consequently, the most well-described sRNA bind indeed in front of the start codon. On the other hand, sRNA target prediction analysis are often interpreted in terms of the best binding energy alone. Considering also the energy needed to make the binding site accessible, systematically favors to predict

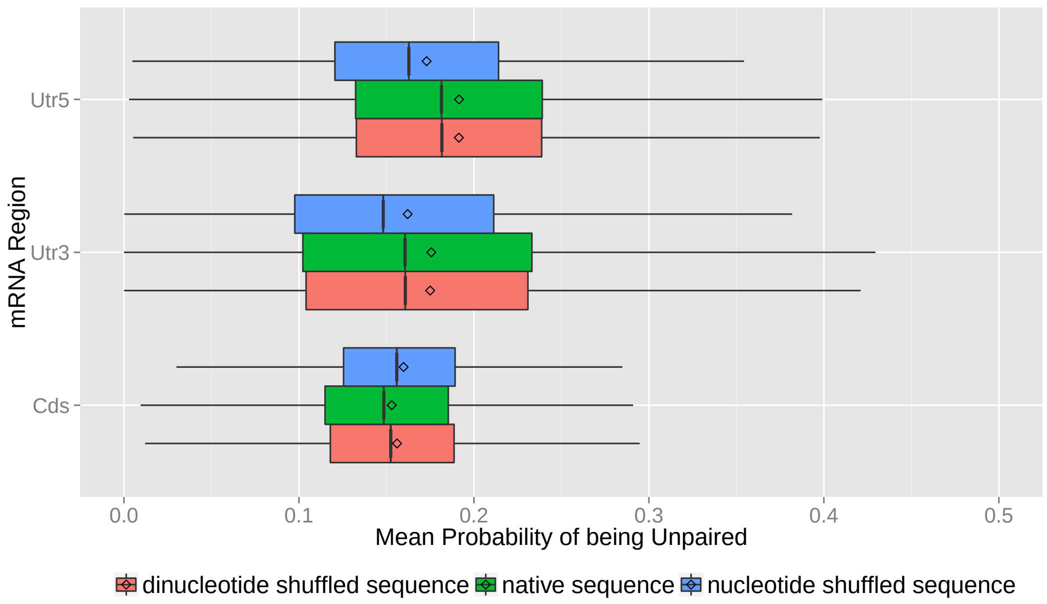


Figure 4.2: From a total of 3,251 chromosomes (all chromosomes in the NCBI database [146] with more than 100 annotated genes) all 7,797,873 annotated genes were used to determine the nucleotide accessibility of different mRNA regions. Therefore, the mean accessibility was calculated for the 5' UTR (30 nt upstream of the start codon), CDS and 3' UTR (30 nt downstream from the stop codon) using `RNAplfold` [17] (parameters: -W 200 -L 150 -u 4). The boxplot shows the distribution of the calculated mean accessibility per gene, compared to the accessibility of a random nucleotide and dinucleotide shuffled variants of the same sequence. The diamonds mark the cumulative average for 5' UTR, 3' UTR and CDS. For readability outliers were omitted from the plot.

sRNA–mRNA interactions around the translation start. It is up to clarification, whether the introduction of this bias is in fact backed-up by biological significance or if it wrongfully neglects other binding sites. Therefore, new techniques to assess how functionally relevant predicted sRNA–mRNA interactions are, will be useful (see chapter 9).

A similar picture can be observed at the end of the ORF. Here, the stop codon is the most accessible position. The following 3' UTR shows a less pronounced pattern than the 5' UTR although the base level of accessibility is increased relative to the base level within the coding region and relative to a randomly shuffled sequence (Fig. 4.2). If this is biological functional or simply an artifact from the following downstream gene, which in bacteria is typically rather close, is at this point mere guesswork since there is no robust knowledge on the extend of the 3' UTR.

The role of accessibility and RNA structure within the coding region is less well examined. A stable secondary structure of the CDS is somehow counter intuitive since it might hinder or stall the ribosome in the course of translation. Nonetheless, several bacterial species, yeast and human show a significant bias in favor of stable local structures in the CDS ([197, 110]. Fig 4.2 bottom lane shows a slight but significant[3] smaller mean accessibility, indicating more stable structure, compared to random shuffled sequences. In contrast to the UTR the accessibility difference and its significance remain if the native sequence is compared to a dinucleotide shuffled sequence. It seems that not only the dinucleotide content but also the sequence itself is selected to retain more structure and less accessibility than expected by chance for a random sequence with the same dinucleotide content. It was speculated that mRNA structure plays a role in RNA processing, regulation of mRNA stability, and translational control [110].

Table 4.2: List of species and their optimal growth temperature (OGT) according to literature.

Species	OGT [°C]	Reference
Colwellia psychrerythraea 34H	0	[220]
Psychromonas ingrahamii 37	15	[220]
Aeromonas hydrophila ATCC 7966	28	[220]
Escherichia coli K12	37	[29]
Chlorobium tepidum TLS	47	[115]
Deinococcus geothermalis DSM 11300	47	[220]
Symbiobacterium thermophilum IAM14863	60	[130]
Thermus thermophilus HB8	70	[220]

Optimal growth temperature and mRNA structure There seems also to be a relationship between the codon usage and the accessibility of the coding region [197]. Fig. 4.3 shows for the examined data set (Tab. 4.2) that the accessibility of the mRNA across different species and growth temperature is less dependent on the temperature than expected by chance. Interestingly, this seems to be a feature mediated by the codon usage. The randomly, synonymously mutated sequences show the same temperature trend than the random sequence, whereas the mutated sequence with the preserved codon usage shows the same trend than the native sequence.

One could reason that if it is advantageous to keep the accessibility for the CDS in a certain corridor, the codon usage is one way to achieve this. Since the structure formation depends, beside other factors, on the temperature, different bacteria with different optimal growth temperature (see table 4.2) adapt their sequences to stay within this corridor. The CDS sequence is under different selection pressures, the most prominent is the amino acid sequence it codes for. An other selection pressure, one could argue, seems to be the maintenance of a favorable accessibility pattern. Here, the redundant genetic code comes in handy. To adjust the mRNA structure synonymous mutation provide some flexibility, to change the RNA sequence without

---

[3] A one sided t-test results in the highly significant p-value of $\leq 2.2 \cdot 10^{-16}$ that the mean nucleotide accessibility is lower in the sample of the native sequences than in the shuffled sequences sample.

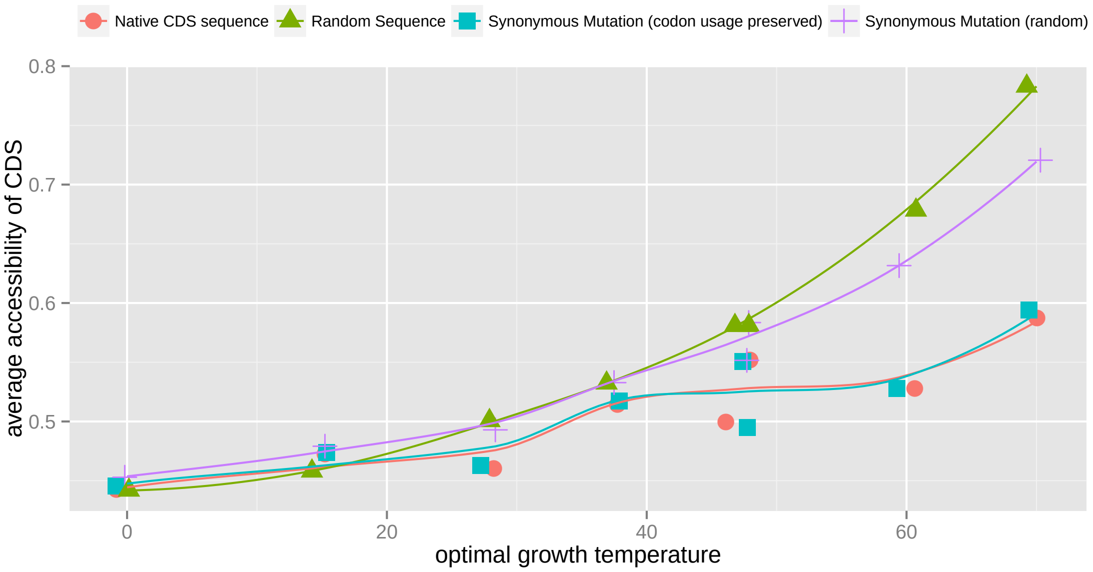


Figure 4.3: Relation between the mean RNA accessibility of the coding region for different species. For each species the sequences were computationally folded at their optimal growth temperature (see table 4.2). Beside the native CDS (red) the accessibility of mutated sequences is also plotted. First, each codon in the CDS was mutated in a way to preserve its encoded amino acid sequence and the overall codon usage as deduced from the whole genome (blue). Second, the CDS was mutated again but this time a random codon with the same amino acid coding was inserted, hence the genes would still code the same proteins but the codon usage is changed to be uniformly distributed (purple). Finally, the CDS was replaced with a random sequence of the same length (green).

changing the encoded amino acid sequence. This might explain that different bacteria show differences in codon usage, which is mostly far away from the randomly expected codon usage [201]. Respecting a given codon usage freezes also $^2/_3$ of the dinucleotide content, which is known to play an important role in RNA structure formation [249]. All this can be seen in Fig. 4.3 , the randomly expected accessibility increases very fast with rising temperature. The calculated mean CDS accessibility for different bacteria and their native sequence stays comparably stable. Even if each CDS codon is substituted randomly with a synonymous codon, respecting the observed codon usage for each species, the accessibility does not change much. On the other hand, if the codons are randomly mutated to synonymous codons with no respect to the codon usage, the calculated accessibilities draw much closer to the randomly expected accessibilities. To the knowledge of the author, this aspect of codon usage was never considered for biotechnological applications. It might be exploit to optimize the design of exogenous expressed genes to the new host.

4.1.6 3' untranslated region

Little is known about the function of 3' UTR is bacteria. One noticeable exception is the GadX/GadW/gadY system from *E. coli.* GadX and GadW are HTH-type transcriptional regulators, which regulate gene expression in respond to acid stress. Both are transcribed together into one polycistronic mRNA. On the opposite strand the gene for small RNA gadY is encoded. Overexpression of gadY leads to a significant enrichment of GadX and GadW protein [166]. Experiments appear to indicate that the hybridization of gadY to the 3' UTR of gadX promotes the cleavage of the gadXW mRNA. On his own, gadX and gadW mRNA seems to be more stable than the longer gadXW mRNA [27].

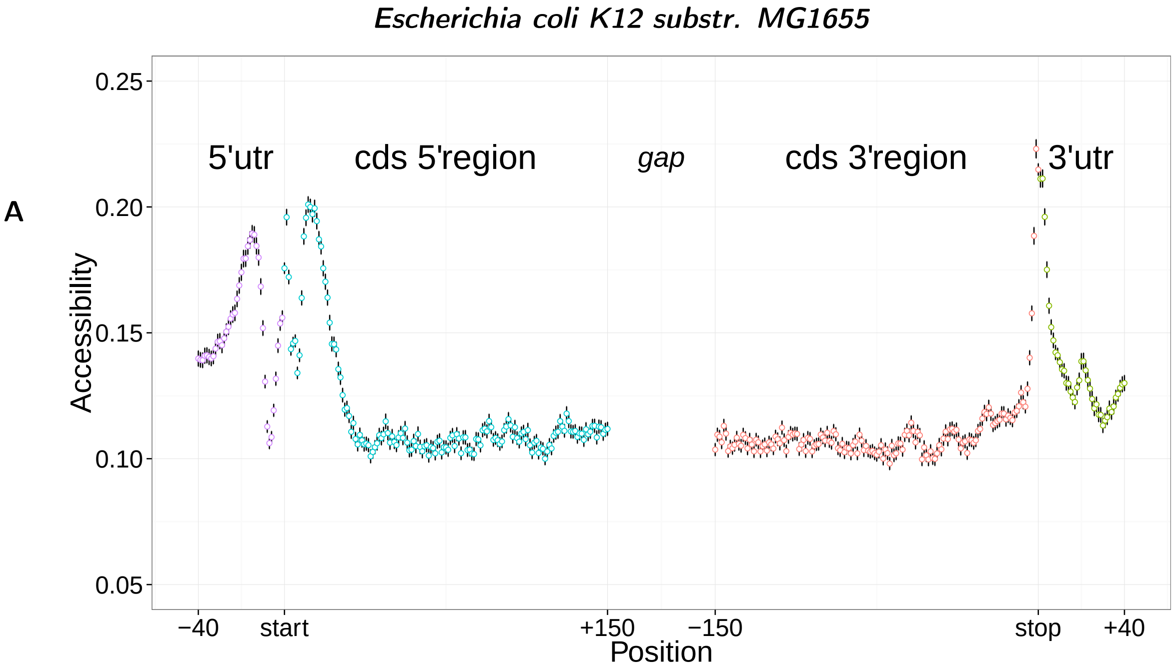
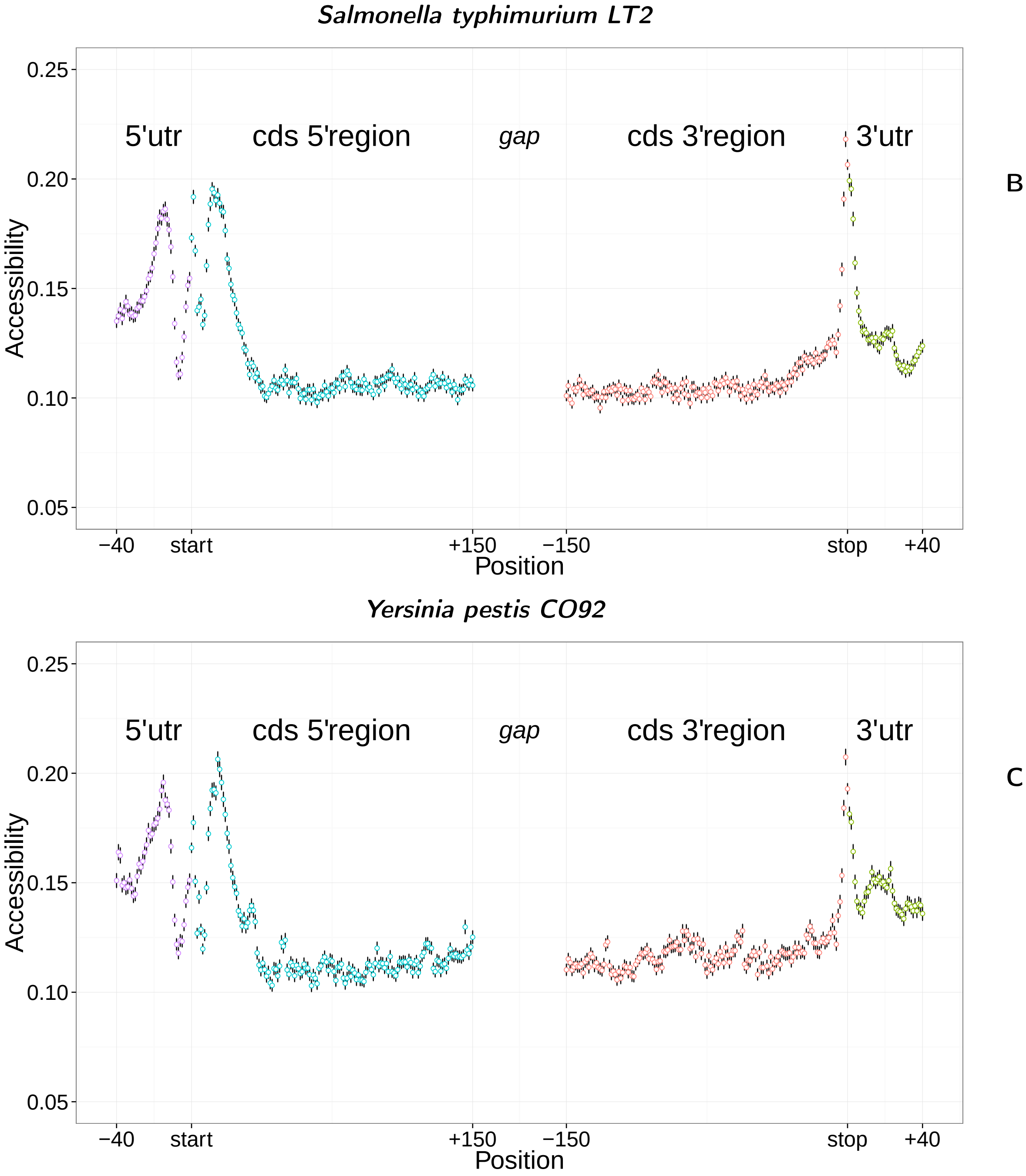
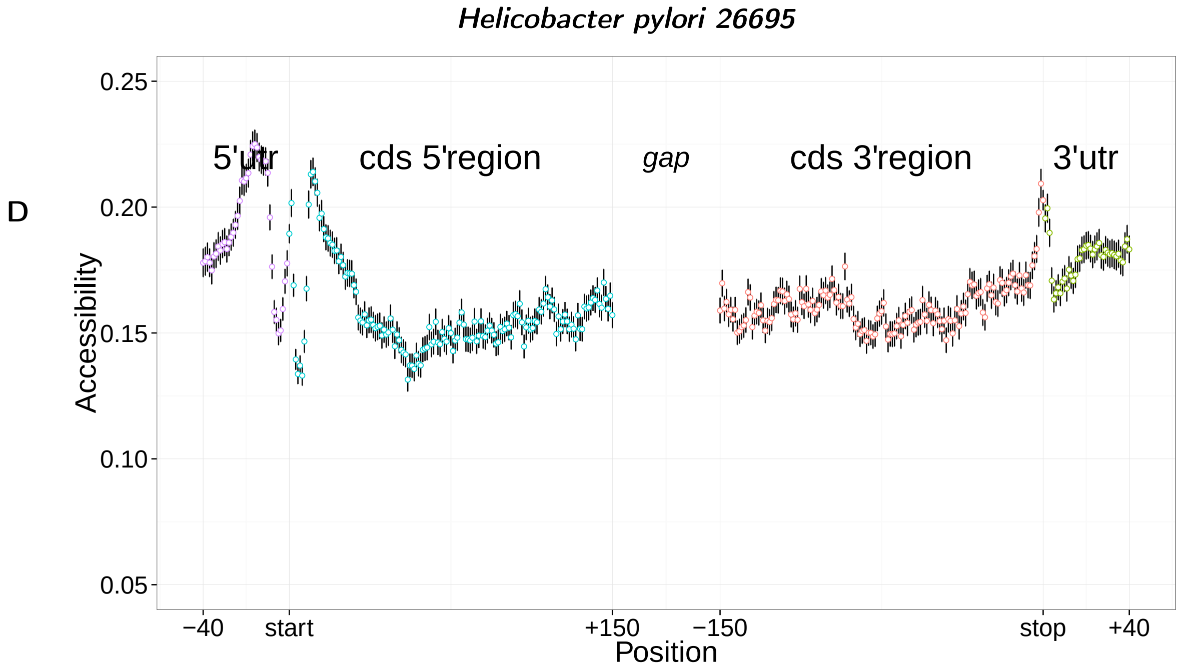
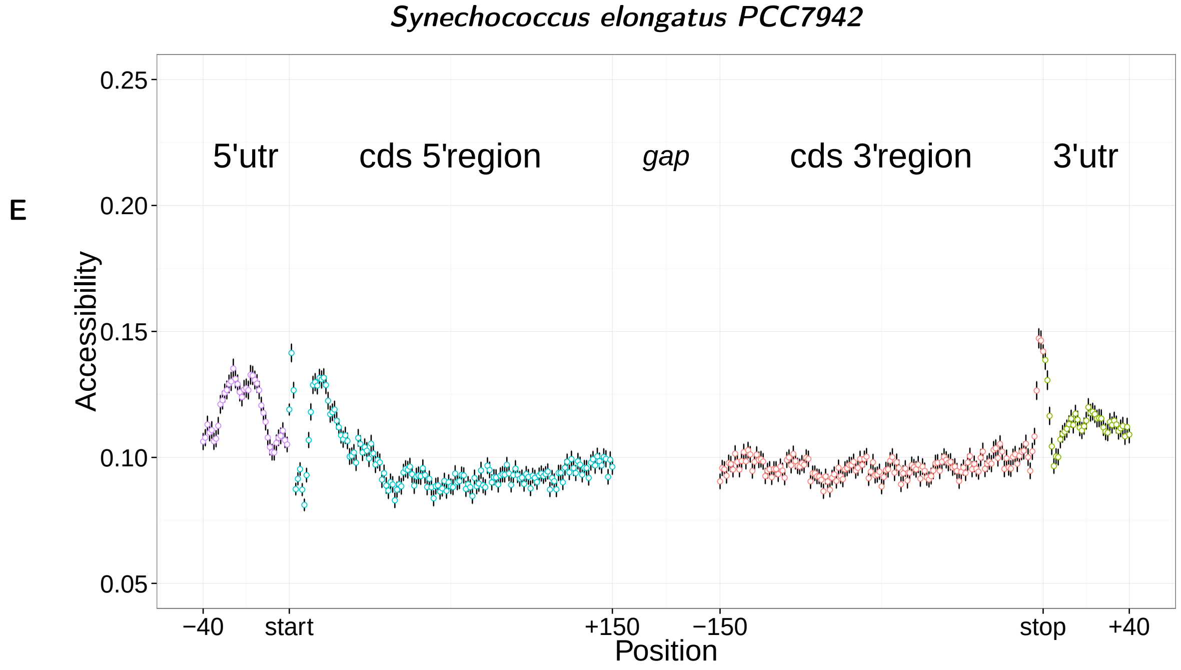
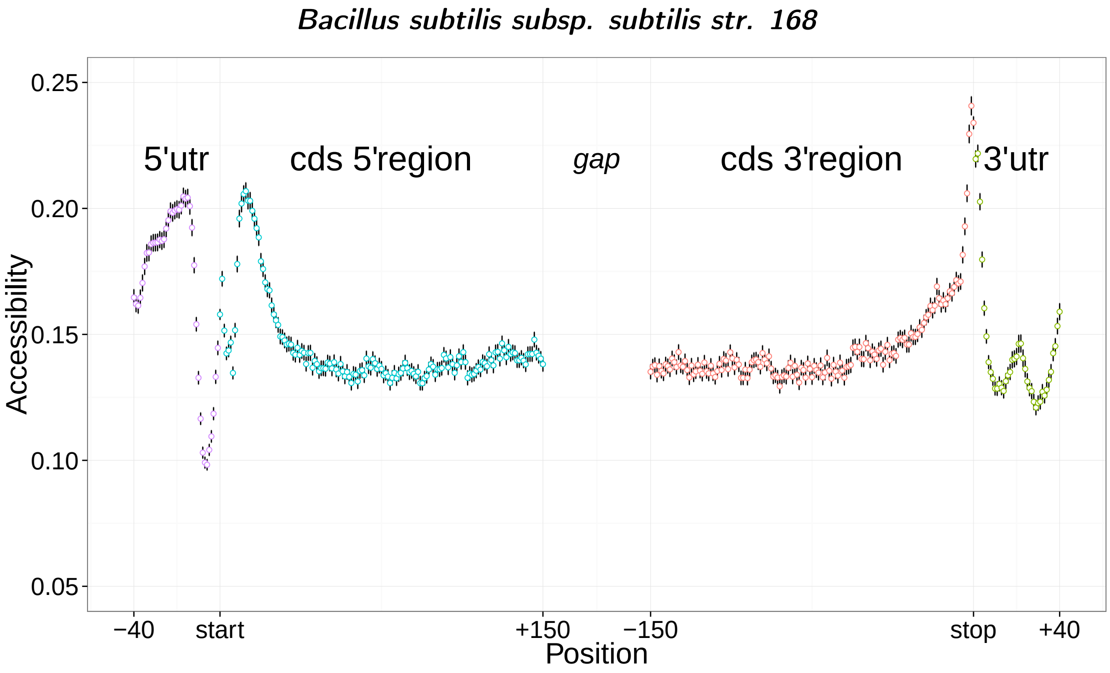


Figure 4.4: The accessibility of the 5' and 3' UTR and the adjacent coding region. For the exemplary chosen species **A** *E. coli* (γ-Proteobacter, Gram-negative), **B** *S. typhimurium* (γ-Proteobacter, Gram-negative), **C** *Y. pestis* (γ-Proteobacter, Gram-negative), **D** *H. pylori* (ϵ-Proteobacter, Gram-negative), **E** *S. elongatus* (Cyanobacteria, Gram-negative), **F** *B. subtilis* (Firmicutes, Gram-positive), and **G** as a control, all above used species are pooled and each sequence was dinucleotide shuffled (using a reimplementation of [101]).
Each annotated gene in the NCBI genbank [146] was used to calculate the accessibility of the coding region with additional 100 nt upstream and downstream using `RNAplfold` [17] with the parameter `-W 200 -L 150`. For each position around the start and stop codon (40 nt from the UTR and 150 nt from the CDS), the mean probability that a region of 5 nt length is unpaired is plotted. The error bars indicate the standard error for each position.
See also subfigures **B** to **G** on page 36-38.


Salmonella typhimurium LT2
5'utr
cds 5'region
gap
cds 3'region
3'utr
B
Accessibility
0.25
0.20
0.15
0.10
0.05
−40
start
+150
−150
stop
+40
Position
Yersinia pestis CO92
5'utr
cds 5'region
gap
cds 3'region
3'utr
C
Accessibility
0.25
0.20
0.15
0.10
0.05
−40
start
+150
−150
stop
+40
Position

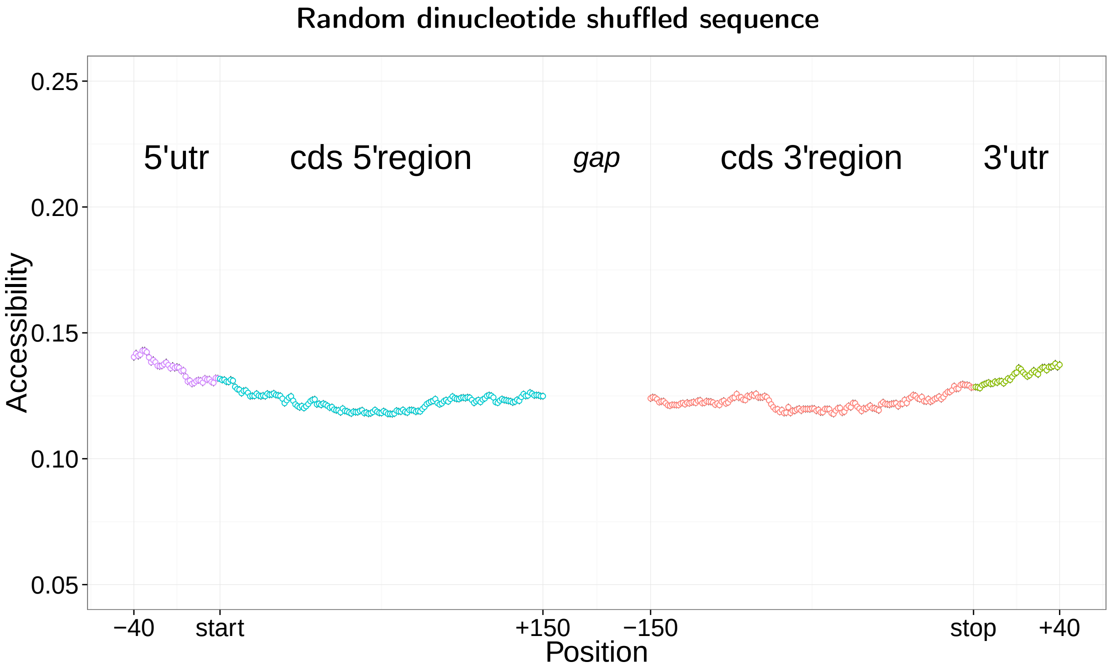

Helicobacter pylori 26695
D
5'utr
cds 5'region
gap
cds 3'region
3'utr
Accessibility
0.25
0.20
0.15
0.10
0.05
−40
start
+150
−150
stop
+40
Position


Synechococcus elongatus PCC7942
E
5'utr
cds 5'region
gap
cds 3'region
3'utr
Accessibility
0.25
0.20
0.15
0.10
0.05
−40
start
+150
−150
stop
+40
Position



Bacillus subtilis subsp. subtilis str. 168

F



Random dinucleotide shuffled sequence

G

4.2 Gene regulation

Bacteria cells show a fascinating broad spectrum of habitats where they can survive. To achieve this diversity they must have developed the ability to adapt their internal program according to signals from the environment. To connect the numerous sensors and receptor with the effectors a fine-meshed information network is spun through the cell. `RegulonDB` [67] coined the term Gensor Units (genetic sensory response unit) for a tightly connected part of these network from the initiating stimulus, over the signal transduction to the change of gene activity and eventually to the cellular response resulting from the modified gene expression.

In the following, we will focus on the step where the transduced signal directly effects the expression of its target genes. Here, two main modes have to be distinguished. First, the gene expression can be activated on the transcriptional level, enhancing the rate of mRNA formation from the genomic DNA. Second, the rate of translation from the pre-produced mRNA can be altered, either by changing the mRNA stability, thus making it longer or shorter time available as a blueprint for protein production, or by increasing/decreasing the frequency of translation initiation events.

4.2.1 Transcriptional regulation

Transcription is the process by which the RNA polymerase produces an RNA copy of a DNA coded gene. This reaction can be divided in three distinct phases: transcription initiation, elongation and termination. In the first stage, transcription initiation, the RNA polymerase binds the double stranded DNA at a defined region, called the promoter, where it forms a bubble in the double stranded DNA, so that the template strand becomes accessible for base-pairing with the transcript to-be. In the elongation phase the polymerase complex moves along the DNA, unwinding it and adding new ribonucleotides to the 3' end of the RNA, forming an RNA/DNA hybrid. When the RNA polymerase reaches the so called terminator, the elongation complex halts transcription. The elongation complex is destabilized and releases the transcribed RNA. This last stage is called termination. Two types of bacterial transcription terminators are known, the rho-dependent and the rho-independent (also known as intrinsic terminator).

Termination Rho-independent terminator is characterized by a GC-rich stretch on the coding strand, hence within the RNA, with the capability to form a stable hair-pin structure, which is followed by a stretch of T bases. Rho-independent terminators can be reliably predicted computational, due to their well defined conservation pattern [113].

For rho-dependent terminators the situation differs. First of all it requires an additional protein factor rho, a hexameric ATP-dependent helicase. Bioinformatic prediction of rho binding sites, which are called rut sites (rho utilization site), is more difficult compared to the rho-independent terminator. The binding to the RNA takes place on an unstructured C-rich sequence upstream of the termination site. A consensus sequence is not required for termination [173]. Chromatin immuno-precipitation and micro-array (ChIP-chip) study in *E. coli* could annotated ∼200 rho-dependent terminators of which seven could be associated to ncRNA genes [172].

Transcription initiation In bacteria there is only one kind of RNA polymerase, in contrast to eukaryotic cells where rRNA, tRNA and mRNA are transcribed from different polymerases. The bacterial holoenzyme, consisting of six subunits ($\alpha_2\beta\beta'\omega\sigma$), with a molecular weight of ∼460 kD. Due to electrostatic properties the RNA polymerase has a general DNA binding affinity. However, the sequence specificity to bind promoters close to transcription start sites (TSS) is provided by the σ factor bound to the polymerase core enzyme. Once the polymerase binds a promoter the closed complex is turned into an open complex by melting a small stretch of DNA. Ribonucleotide polymerization can start, generally this happens first without movement of the holoenzyme. This way small pieces of RNA from a few up to 20 nt length are produced, which are called abortive initiation products. Transcription proceeds into elongation phase by a process called promoter clearance. The main event in this process is the dissociation of the σ factor from the holoenzyme leaving the core enzyme to elongate along the genomic DNA, producing a functional transcript [120].

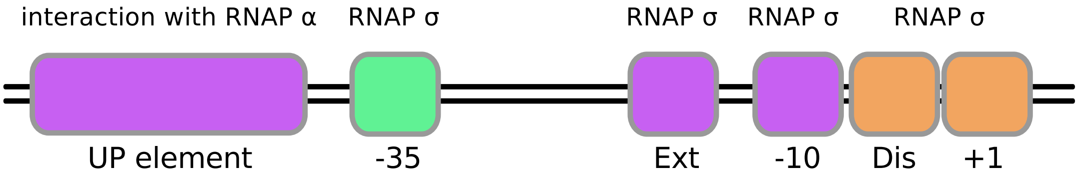


Figure 4.5: A schematic illustration of the bacterial promoter structure. (Adapted from [120])

Promoters DNA sequences in the genome upstream of a gene with the ability of specific RNAP binding are called promoters. They enable a regulated gene expression. In bacteria, promoters consist of several elements which are to different degrees obligatory for efficient transcription (see Fig. 4.5). Most importantly is the so called −10 element, named due to its position around 10 nt upstream of the transcription start site. This element has a consensus sequence of TATAAT, hence its alternative name TATA-box. This region makes substantial contact with the σ factor. It is also the site where the double stranded genomic DNA is opened, forming a bubble which gives access to the inward pointing bases. The −35 element interacts also with the σ factor. Its consensus structure is TTGACA. The region between the −10 and −35 element is not very specific in its sequence but the distance between them has substantial influence on transcription efficiency. Further upstream of the −35 element the UP element can be found.

This one is somewhat special since it interacts with the α subunit of the RNAP holoenzyme. Further sequences around transcription start site has been shown to interact with the σ factor but are less conserved. Here the extended -10 element and the region between the -10 box and the TSS have to be mentioned. Although they are not well conserved they can still influences the stability of the open RNAP DNA complex, thus can have an effect on transcription rate [120].

Transcription factors `EcoCyc`, an integrative scientific database for the bacterium *Escherichia coli K-12 MG1655*, classifies ∼4 % of the *E. coli* genes as so called transcription factors (175 of the 4489 annotated genes [112]). Transcription factors are proteins with the ability to alter the transcription of their target genes in a controlled fashion. Transcription factors are classified in several families according to their two functional domains. One part is responsible to sense a stimulus, this might be a ligand-binding, protein-protein interaction or a phosphorylation. The other part is responsible for binding the DNA and subsequently influencing the interplay between the RNAP and the promoter [11].

A special kind of bacterial transcription factors are the afore mentioned σ factors. These factors are essential for correct promoter recognition by the RNAP. With this mechanism it is possible to regulate whole classes of genes globally. For example the σ_{38} factor (RpoS) regulates more than 70 genes involved in stress response [88].

ncRNA and transcription regulation In eukaryotic cells many examples of ncRNA influencing transcription are known [71]. The mechanism range from altering chromatin modifications[4], via modulation of activator functions[5], to direct control of the eukaryotic transcription complex and its activity[6].

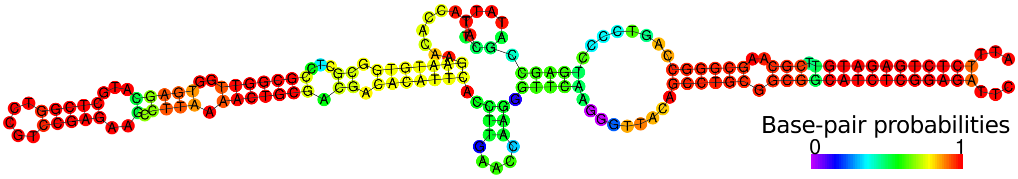


Figure 4.6: Centroid secondary structure of *E. coli*'s 6S RNA (gene symbol: ssrS). Sequence obtained from NCBI genbank [146]. Structure calculated and drawn with `RNAfold web server` [76, 91]. Colors code for the probability that a base in a base-pair is paired or that a base in an unstructured region is unpaired, respectively.

---

[4] HOTAIR would be one prominent example. It increases trimethylation of histon H3 K27 within the HOXD locus, thus decreases transcription rate [186]

[5] For example HSR1 (heat shock RNA 1) together with eEF1A leads to trimerization of HSP1 (heat shock factor 1) upon temperature stress. Only the trimeric HSP1 has DNA binding properties and can activate its target genes [198]

[6] Human ALU RNA binds to eukaryotic RNA polymerase II in response to heat shock and subsequently represses transcription [138]

In bacteria there is less known about regulatory function of ncRNA acting at the level of transcription. One well established example is the role of 6S RNA in the cellular response to stress due to lack of nutrients. *E. coli*'s 6S RNA was shown to be a 183 nt long RNA which forms an elongated stem with unpaired internal bulges (see Fig. 4.6). This structure mimics the DNA of a promoter inside the open complex during transcription initiation. This enables the RNAP associated with the σ factors σ_{70} or σ_S to bind 6S RNA which leads to a competition between functional promoters and 6S RNA, reducing consequential translation initiation events due to a decreased number of available RNAP. This down-regulation can be observed for many but not all σ_{70} dependent promoters. Genes which seems not to be effected have an extended -10 promoter element in common [226].

6S itself is highly up-regulated upon entry into the stationary growth phase. At that time, more than 75 % of RNAP associated with σ_{70} are complexed with 6S RNA, indicating a physiological role of 6S RNA during this phase. However, Δ6S RNA mutants show a reduced vitality not until three weeks of growth in stationary phase, compared to the wild type. This implies a role for 6S RNA in enduring extended phases of nutrients deprivation [226].

Experimental approaches Transcription start sites (TSS) and promoters are co-occurring in the genome. Annotated TSS can be used to find novel features of promoters. Hitherto, the flow of information is often the other way around. Usually, characteristics of promoters are used to detect TSS. This strategy potentially tends to oversee unusual elements which can have an influence on transcription initiation. Novel high throughput techniques, e.g. differential RNA-seq, to annotated original primary TSS were developed to overcome this limitation. In chapter 7 a statistical analysis of such generated data is presented. This might help in the thorough annotation of TSS and subsequently provides new data of putative promoter regions. This region can be analyzed to extract sequence and structure feature which might be important for gene regulation in bacteria.

4.2.2 Translation regulation

Translation is the process by which the information stored in the nucleotide sequence of an mRNA is used to synthesis a protein. The central translation machinery is the ribosome. The fully assembled ribosome has a sedimentation coefficient of 70S and can be further divided into two subunits the small, or 30S, and the large subunit with 50S. Both subunits are RNA - protein complexes. The 16S ribosomal RNA (rRNA) forms together with 21 proteins the small ribosomal subunit. The large subunit consists of the 23S rRNA and 31 proteins [68]. The ribosome has three separate binding sites for tRNA, *A*, *P*, and *E*, which stands for aminoacetyl, peptidyl and exit site, respectively.

Ribosomes have the ability to assemble on an mRNA molecule, recognizing the correct start codon and translation frame, sliding along the mRNA, translating the whole protein coding region, and disassemble at the correct stop position to free the newly formed protein and the reusable mRNA.

All steps in the course of translation, i.e. initiation, elongation and termination, are regulated by a number of different factors. Gene regulation on the post-transcriptional level contributes as much as three orders of magnitude to the overall variability of gene expression [122]. In the following section, the basic events for each step will be described with emphasis on how they are subject to regulation.

Translation initiation In bacteria translation initiation occurred already co-transcriptional, i.e. during transcription, when the 3' end of the mRNA is still synthesized, the ribosomal machinery can already assemble on the 5' end of the mRNA [124]. The initiation is prepared by the binding of the initiation factor IF3 to the 70S ribosome, which leads to the dissociation of the 70S into the 30S and 50S subunits [174]. IF1 follows, by binding into the A-site of the 30S ribosomal subunit enhancing the effect of IF3 [78] and blocking the initiator tRNA from entering the A-site [47].

After the subunit dissociated, the mRNA binds the ribosome, immediately followed by IF2 and initiation tRNA fMet-tRNA$_f^{Met}$. This forms the relatively unstable 30S pre-initiation complex [124]. The mRNA interacts with the Shine-Dalgarno sequence (SD), which is normally located upstream, near the translation start site on the mRNA. The SD interacts with the so called anti-Shine-Dalgarno sequence, which is located at the 3' end of the 16S rRNA. The fMet-tRNA$_f^{Met}$ binds into the P-site on the 30S ribosome and interacts also with the IF2. After IF1 and IF3 dissociate and the anti-codon loop of the initiator fMet-tRNA$_f^{Met}$ interacts with the start-codon on the mRNA. From here on, the reading frame of the mRNA is determined. The dissociation of the last initiation factor IF2 promotes the joining of the 50 S ribosome to build up the more stable 70S initiation complex. At this stage, the ribosome is ready to enter elongation phase, and to polymerize amino acids, according to the blueprint encoded in the mRNA, to produce a polypeptide eventually.

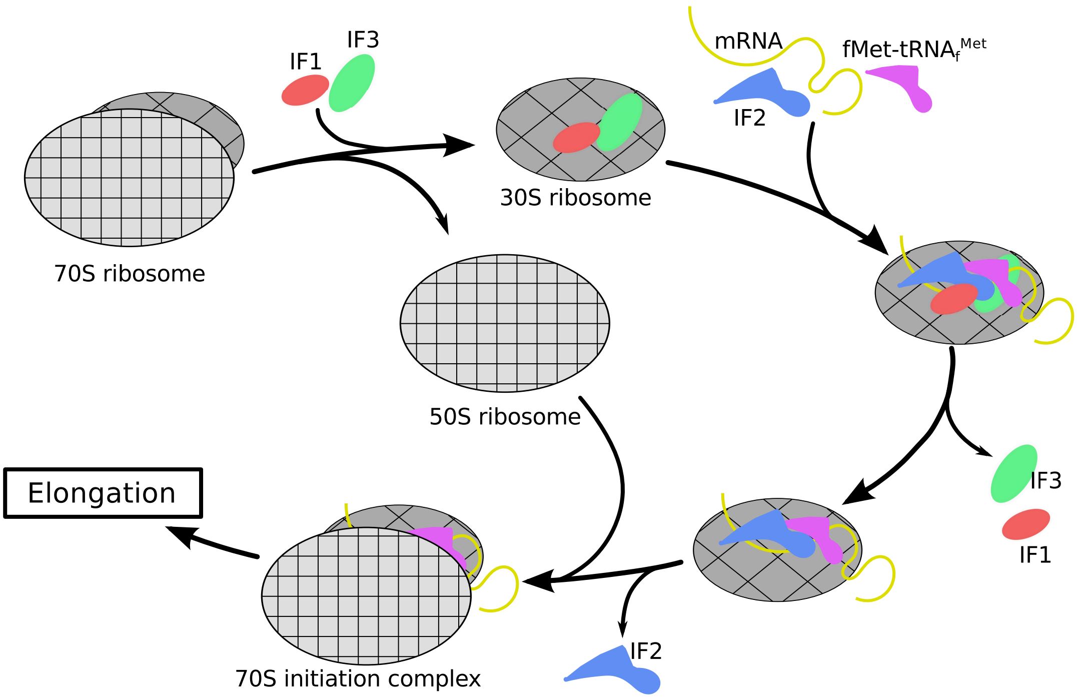


Figure 4.7: A schematic illustration of the bacterial translation initiation process. The route from the unbound 70S ribosome, via the pre-initiation complex to the complete 70S initiation complex, which leads further on into the elongation phase, is depicted [124].

Regulation of translation initiation

Riboswitches are complex non-coding RNA structures which alter gene expression of associated genes upon binding of a ligand-metabolite. On one hand riboswitches can function on a transcriptional level by forming transcription terminators and thus prevent mRNA formation [239]. On the other hand, on a translational level, they can change the structure of the ribosome binding region to ease or hamper the translation initiation [12].

An interesting example of a riboswitch which functions both on the transcriptional and translational level is the so called thi box. This structure, which is well conserved among Gram-positive and Gram-negative organisms, is found in front of genes involved in thiamine (vitamin B_1) metabolism. Binding of thiamine to the thi box structure induces a refolding, causing the SD sequence to be mask by a stable hairpin. This results in inhibition of the ribosome to initiate translation. Without elongating ribosomes on the mRNA a stable transcription terminator hairpin can form a few hundred nucleotides downstream of the translation start, leading to a premature termination of transcription and a non-functional mRNA [152].

RNA thermosensors are in some instances similar to riboswitches. Since all biological processes are temperature dependent, a thigh control of gene expression due to temperature changes is needed. One very obvious application of thermosensors are the control of virulence genes. Many pathogens can survive in the environment. The first indication that they entered a potential host is often a temperature shift from ambient temperature to the body temperature of their host. To adapt as quick as possible, many virulence genes are translated under the control of such RNA thermometer.

RNA thermometers are complex structures in the 5' UTR which partially overlap with the ribosome binding site of their supervised gene. Already small temperature changes (in the range of ∼1°C) can lead to a structural rearrangement and an exposition of the SD sequence [118].

Regulation by competition An exceptional mechanism of post-transcriptional gene regulation was discovered with the threonyl-tRNA gene. In this particular case, the gene product acts in a negative feedback loop on its own production. In more detail, threonyl-tRNA synthetase binds the translation initiation region (TIR) of its own mRNA under conditions of high threonyl-tRNA synthetase concentrations and hence represses its own expression [213].

small RNA In its mechanistic details unique to prokaryotes are small RNA which can can bind different target mRNA and alter their translation or stability. A thorough introduction is given in the chapter 5.

Translational coupling of genes within polycistronic mRNA In contrast to eukaryotes, bacteria genes are organized into operons, which adds an other layer of complexity and potential for regulation. Polycistronic mRNA often possess more than one open reading frame. Within the transcript these ORF can either overlap with each other, forming particular stop and start codon arrangements [192], or they are separated by an inter-cistronic spacer between two consecutive ORF (see Fig. 4.1).

Whether the translation initiation at the downstream genes differs mechanistically from the first ORF is still a debatable question. Current models for translation initiation of the second ORF assume either disassembly of post-termination complex and de novo initiation at the second cistron translation initiation region. Hence, the same processes guide upstream and downstream gene translation initiation within one transcript. In contrast, some models assume the migration of the post-termination ribosome or the 30S subunit along the mRNA scanning for a following initiation codon. In this case the two initiation mechanisms would differ. It is also possible that both models could be realized with different frequencies [167].

As pointed out, this question is not yet settled but some experimental results indicate that at least the translation rate of the first gene has an effect on the translation rate of the second gene. Further on, this effect is dependent on the distance between the two genes. This results hint at least to favor the second model, the recycling of the partially assembled ribosome. Translational reinitiation at a distance after the stop codon of the preceding ORF implies that the assembled ribosome slides down along the mRNA [1]. The ribosome sliding might be hindered or decelerated by inter- or intra-molecular secondary structure elements in the intercistronic spacer [102]. This model is also able to explain the observation that some polycistronic mRNA show different protein synthesis rates for the different genes. This phenomena has been termed discoordinated expression [156]. For the gal operon, coding for different genes involved in the galactose metabolism, the relative synthesis of the enzymes UDP-galactose-4-epimerase (GalE), galactose-1-phosphate uridyl transferase (GalT), and galactokinase (GalK) vary under different growth conditions.

It was suggested that transcription termination at intercistronic regions, preferential degradation of the promoter distal portion of the mRNA, or different translational efficiencies of the two gal transcripts[7] could explain the discoordination ([228, 105, 179]. Interestingly, the differentially discoordinated expression of the gal operon seems to be regulated by the small RNA Spot42 [156].

mRNA decay To change the amount of newly produced protein, the translation rate or the translation substrate concentration can be altered. The later is achieved either by changing the mRNA production or the mRNA degradation. Therefore, it is straight forward to assume that also mRNA decay is regulated. On average *Escherichia coli* mRNA have a half-life of 3.69±0.49 minutes [20], but the individual mRNA half-life can differ by two orders of magnitude [16]. To intervene with mRNA stability, bacteria employ a diverse set of endo- and exonucleases with distinct functionality.

In *E. coli* the orchestrated degradation of mRNA molecules generally follows a pattern. First, the mRNA is endonucleolyticly cleaved. Afterwards, the fragments are exonucleolyticly digested in an 3' → 5' direction [162].

The stability of mRNA is linked to the overall translational activity [116]. This has the convenient side effect that many post-transcriptional gene regulation mechanisms are also reflected in a change in mRNA abundance, hence, can be studied by transcriptomic- instead of proteomic-techniques[8], because e.g. ncRNA mediated decrease in translation rate leads to a reduced mRNA

---

[7] The polycistronic gal transcript is transcribed from two different promoters, producing two transcripts differing in the transcription start site and hence their 5' UTR [156].

[8] In fact, mRNA abundance and protein abundance are in general not too strongly correlated. For human cell lines it could be shown that only 40 % of the protein level can be explained by the abundance of the corresponding mRNA [216].

half-life, and, as a consequence, a measurable change in mRNA abundance.

Additionally, there are several examples known, where ncRNA directly modify mRNA stability directly. In *Staphylococcus aureus* for example, the RNAIII specifically base pairs with the virulence gene spa. This short RNA duplex is recognized by the endoribonuclease III (RNase III) and subsequently degraded [99].

If such processing sites are conserved new sequencing based techniques might be employed to detect them. One, for this purpose so far not tested method is dRNA-seq [250, 200]. The methodology presented in chapter 7 might be applied for this in a very straight forward manner (see chapter Discussion page 131 ff).

5 Small RNA

Bacterial small RNA are an abundant class of trans-encoded regulatory RNA. They are typically between 40 and 200 nt long [203], with some exceptions which are up to 500 nt long [96]. They have in common that once they are transcribed, they diffuse through the cytoplasm and interact with their targets. This distinguishes them from riboswitches which are physically attached to their gene under control, and act only in this well defined configuration. Sometimes sRNA are subdivided into *cis* and *trans* acting sRNA. The first implicates that the sRNA gene lies on the opposite strand of its target gene, having, as a consequence, a long and perfect complementarity to the target mRNA. In contrast, trans-acting sRNA, are encoded apart from their targets, and have only a very short and imperfect complementarity to their targets [126]. They have the ability to base pair with a short and imperfect complementarity to the mRNA of their target genes. It has to be emphasized that here the plural “genes” is used with good cause since there is increasing evidence that sRNA having just one target is the exception rather than the rule [4]. Multiplicity is also possible the other way around. Many mRNA are targeted by more than one sRNA, serving as a hub, connecting the sRNA mediated regulatory network with other signaling network [22].

Due to the fact that sRNA act on the translational level and by a different mechanism than transcription factors (TF, see section 4.2 and 5.1) they show very different kinetic behavior in the regulatory network. Simulations have shown that TF based networks are better suited for quantitative signaling, whereas sRNA based circuits perform better in qualitative signaling. This allows the cell to change rapidly between different states in response to large changes in input signal [148]. This theoretical discovery is in perfect agreement with the biological role of already well studied sRNA in the bacterial cell. It has been reported that small RNA are involved in the regulation of important biological processes, such as virulence, stress response and quorum sensing [15, 142, 224].

5.1 Mode of action

Trans acting sRNA can be subdivided into two major category, defined by the type of targets they interact with.

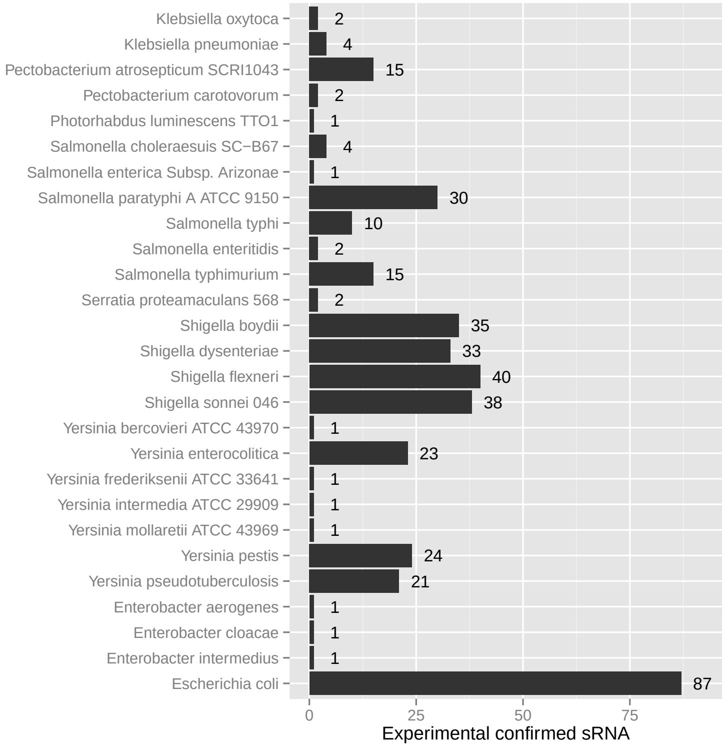


Figure 5.1: Number of experimentally verified sRNA for different species [96].

5.1.1 Regulation of protein activity

In *E. coli* there are three sRNA known which interact directly with a protein to modulate their activity. One example is the already mentioned 6S RNA (see page 41). The other two are CsrB and CsrC which both bind to the protein CsrA. This protein regulates several genes involved in carbon storage by binding their 5' UTR and inhibiting their translation. Both sRNA, CsrB and CsrC, share a common sequence motif with the targets of CsrA. After binding of one of the sRNA to CsrA, the protein is no longer available to inhibit the translation of its targets, which are as a consequence translated with a higher rate [190].

5.1.2 Regulation of cis encoded mRNA activity

Some sRNA target directly the mRNA of the gene encoded on the opposite strand at the same genomic locus. This so called anti-sense sRNA are well studied from plasmid toxin-antitoxin systems. The basic principle is the different stability of sRNA and mRNA from the same loci. One example is the hok/sok system. Hok is a lethal protein with a stable mRNA. Encoded anti-sense to hok, the unstable sok RNA is transcribed. Both genes originate from a plasmid. As long as the plasmid is present in the cell, the sRNA and mRNA are constantly produced.

Sok prevents the translation of the toxic hok protein, hence the cell can proliferate. Once the plasmid is lost, the concentration of the unstable sok drops faster than the stable hok. After some time the remaining sok sRNA can no more block all hok mRNA and the lethal toxin protein is produced. This way the plasmid secures its own propagation in the bacterial population [69].

5.1.3 Regulation of trans encoded mRNA activity

The predominant and also more interesting system works by sRNA which interact with distantly encoded, i.e. trans encoded, mRNA to alter their translation frequency. Although there are an increasing number of studies showing that the interaction can also take place in the coding region of the mRNA [175, 252, 236], in general, the sRNA interacts with the 5' UTR of the mRNA. The base pairing region is normally rather short[1] and imperfect. In some cases the sRNA and the mRNA interact with two separate very short sequence stretches. OxyS is a well known example forming such a *kissing hairpins*. This particular sRNA interacts with its target fhlA, a transcriptional activator, via two complementary regions which are on the sRNA 67 nt apart[2]. The contact is established via seven and nine base pairs, respectively [7]. As we will see later on, this cases are a especially problematic in matters of computational target prediction (see section 5.3.2).

Once the sRNA has bound its target mRNA, a series of events are triggered, leading to an altered rate of protein synthesis. Here, the most fundamental classification is the net effect of the regulation, whether it leads to an increase or a decrease of protein synthesis from the particular mRNA target.

Negative regulation In contrast to comparable mechanisms in eukaryotic cells there is no dedicated protein machinery which simply applies sRNA for specific target detection, such as the miRNA system with its RNA-induced silencing complex (RISC) [87]. In general, targets of bacterial sRNA are translationally silenced by blocking the access of the ribosome to the ribosome binding site (RBS) [244] (see figure 5.2A). This prevents translation initiation. Subsequently, the unused, thus not ribosome covered mRNA, is degraded by RNase III or RNase E [140, 23]. This leads to a change in mRNA abundance which can be seen by transcriptome profiling[3], although in some cases no change in the overall transcript stability can be observed [156].

The mRNA stability can also be influenced directly by sRNA. Thereby, the sRNA can bind its target also further downstream, in the CDS or in the intergenic region of polycistronic tran-

---

[1] For *Salmonella*'s RybB it was shown that 7 nt are enough to repress its target [170].

[2] On the mRNA the distance is only 42 nt.

[3] By this, laborious and expensive proteomic studies are generally not necessary, which facilitates working with sRNA tremendously.

scripts. Here, an active recruitment of endonucleases by the sRNA·mRNA complex is involved in complex degradation and eventually results in less mRNA substrate for the production of protein [175].

Direct and indirect effects on mRNA stability can be difficult to distinguish experimentally, since both lead to a decrease in mRNA concentration. Especially since both mechanisms can be in effect simultaneously. For one case, RyhB and its target SodB, it could be shown that at first the sRNA binds the mRNA at its translation start site[4] [234] and is indeed competing with the 30S ribosome. This not just leads to a decrease of translation initiation rate (and accompanied mRNA decay by neglect) but subsequently to a specific recruitment of the RNA degradosome. The mRNA is cut as far as 350 nt downstream of the bound sRNA [178]. One proposed advantage of this two-step mechanism is that already elongating ribosomes can finish translation, although new ribosomes are hindered to start translation. When finally the mRNA is degraded no ribosomes are stalled on top of the transcript which would reduce the pool of available ribosomes [178].

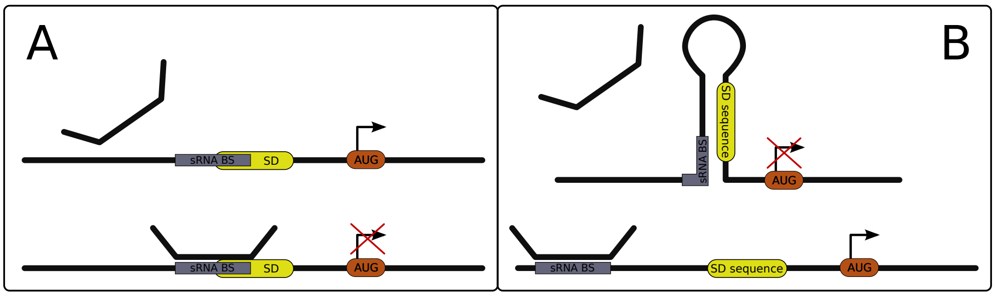


Figure 5.2: Mechanistic scheme of (A) down regulation by blocking the Shine-Dalgarno sequence for the ribosome. (B) Up regulation by liberating the SD sequence from a blocking secondary structure by sRNA binding close-by. Adapted from [64].

Positive regulation The sRNA mediated boost of protein synthesis from a target mRNA is generally induced by dissolving inhibitory structures around the translation initiation region (TIR). Some mRNA show, although abundant in number, only a very low translation rate because the translation initiation efficiency is reduced by adverse secondary structures which prevents the ribosome to recognize the TIR. sRNA can induce a refolding of such structures by binding close-by and competing with the SD sequence for the intra-molecular binding partner (see Figure 5.2B) [64]. The first discovered example of such an anti-antisense regulation was *Staphylococcus aureus* gene hly. An α-toxin which promotes lysis of eukaryotic host cells of the pathogen *S. aureus* [75]. In vitro structure probing and computational analysis indicate that the mRNA region ∼140 nt upstream of the start codon of the hly gene folds back to pair with the SD sequence, blocking it for the ribosome. RNAIII can interfere with this structure

---

[4] The interaction runs from 12 nt in front down to 5 nt behind the start of the translation start [234].

by stably binding the anti-SD sequence releasing the TIR and eventually leading to a dramatic ∼70-fold increase of translation rate [158].

5.1.4 Bi-function sRNA

Although the term small non-coding RNA is still widely used in literature, there are meanwhile a few examples of genes which act as a trans-encoded small RNA on the RNA level, and at the same time code for a small protein in a polyglot manner. SgrS for example inhibits the translation initiation of glucose transporter by blocking the translation initiation region on the mRNA. The same gene also codes for a 43 amino-acid long polypeptide which itself inhibits glucose transporters. [229, 240]

5.1.5 Hfq

Hfq was first identified as a host factor for the RNA phage Q_β, since it was shown that Δhfq mutants are immune against phage Q_β infestation [73]. Meanwhile, the Hfq protein proved to be highly conserved in several bacteria [195]. In *E. coli* Hfq is highly abundant with 30,000 to 60,000 copies per cell [106]. Hfq is a pleiotropically acting RNA-binding protein. This property is mediated through its structure. The monomeric Hfq protein contains a so called Sm-fold, therefore it belongs to the group of Sm-like protein. Sm proteins in eukaryotes are involved in different RNA processing events, such as splicing and mRNA degradation [169]. The Hfq monomers assemble to a hexameric ring, with two separate RNA binding interfaces (Fig. 5.3). Hfq was reported to interact with A-rich motifs on the distal site, and with AU-rich RNA sequence motifs on the proximal site [127]. Furthermore, it was shown that Hfq binds ATP and has ATPase activity, which seems not to be essential for RNA annealing [84].

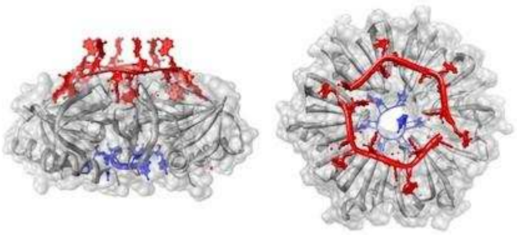


Figure 5.3: Atomic structure of the Hfq protein associated with two RNA molecules on the distal (red chain) and proximal (blue chain) binding site [193, 25].

The important role of Hfq commenced to be recognized after the first Δhfq mutant showed a very severe phenotype: impaired growth rate, altered cell shape and increased sensitivity to UV light. Since this phenotype strongly resembled known rpoS mutants, first guesses proposed that Hfq is involved in the regulation of RpoS expression. This assumption could be shown to be correct [232]. Meanwhile, many more genes, whose regulation is modulated by Hfq, are known. In the most cases, Hfq mediates the formation of the sRNA·mRNA complexes. To do so, Hfq utilizes its two binding sites, one for the mRNA and one for the sRNA. In this sense Hfq serves as an RNA chaperon, facilitating the sRNA-mRNA interaction. For many of these interactions, it was shown that Hfq is essential, for other interactions Hfq increases the regulative effect but is not essential or even does not play a role at all. This observation could be explained by the property of Hfq to recruit RNases. In the presence of Hfq the mRNA gets actively degraded, without Hfq the RBS is blocked by the sRNA which is sufficient to down regulate mRNA translation.

The precise function and role of Hfq is not yet clear and the emerging picture indicates that there is not one role Hfq plays but that it is involved versatile regulatory mechanisms in the different systems. For its role in sRNA based gene regulation the pivotal question which awaits answering is how Hfq gains its RNA binding specificity [81] and if there are other protein factors, which might be involved in the establishment of specificity in the sRNA – mRNA hybridization.

5.2 Finding small RNA

5.2.1 Experimental discovery of sRNA

The first trans encoded sRNA were mostly found by genetic screens. Once a phenotype was observed, mostly by serendipity, the underlying genes were tracked down. DsrA for example, was found in the course of examining the capsular synthesis of *E. coli.* A mucoid phenotype with capsular over-production was associated with a multi-copy plasmid carrying short genomic region downstream of the RcsA gene[5]. This sequence was subsequently identified as an ncRNA, since it has no potential to code for a polypeptide [207]. In the years to come, two targets of DsrA were discovered. First the H-NS transcription regulator, which also causes the phenotype observed in the first place. Soon after the mRNA RpoS could be shown to be not just a target of DsrA, but also to be translational activated upon DsrA binding [135]. Since RpoS has an extraordinary long 5' UTR, it was assumed that it might be targeted by more than just one sRNA. These hypotheses was exhaustively tested by expressing random genomic *E. coli* fragments within an *E. coli* strain with recombinant rpoS::lacZ gene for easier read out. By screening 25,000 colonies a novel sRNA, rprA was discovered [237].

---

[5] Which why this locus was named "downstream from RcsA", or shortly DsrA.

After this initial phase of trailblazer studies the regulatory importance of sRNA in bacteria were acknowledged and genome wide screening projects for new sRNA started. One of the successfully applied methods were genomic tiling arrays [225]. In spite of its successful application, a critical drawback of this method are the high costs. Commercially available micro-arrays cover usually only the already annotated genome, making them useless to detect new transcripts from intergenic regions. Custom arrays could overcome this problem but were and still are very expensive. The biggest advantage of tiling arrays for ncRNA detection is the additional information such an experiment can deliver. Foremost it becomes possible to gain not just potential sRNA loci but also their transcriptional profile, which is already a first step beyond identification towards characterization of the gene [237].

The next great technological leap was the propagation of low cost deep RNA sequencing approaches. With this technique it becomes possible to combine the advantages of micro-arrays with reduced cost and easy automation of the screening procedure.

Once the special relationship between the RNA chaperon Hfq and sRNA was recognized, Hfq was utilized to discover new potential sRNA candidates. For that, a cell lysat with total RNA is incubated with tagged Hfq. After co-immunoprecipitation of Hfq with attached RNA, the nucleotides are sequenced or identified by tiling arrays. In *Escherichia coli* [254] and *Salmonella enterica* [206], this approach was successfully applied to expand our knowledge of sRNA. In *Salmonella*, for example, the number of annotated sRNA was doubled to 64, revealing the important role of sRNA, and Hfq, for the pathogens, in establishing an infection.[206].

5.2.2 Computational discovery of ncRNA

Since sRNA are not translated and thus are not characterized by an ORF, common gene annotation strategies fail to include them in their gene prediction [126]. That is why a whole arsenal of in silico methods were developed to make up for this short coming. So far, all of them have their specific advantages and disadvantages and it is still essential to combine different strategies to obtain the whole picture of the diverse sRNA in a particular species.

Two main characteristics of ncRNA can be exploited to annotate de novo sRNA in a genome. On one hand, ncRNA are, as any functional genomic entity, expected to be conserved over some evolutionary time. On the other hand, ncRNA often have a characteristic structure, whose signature can be detected in a less structured surrounding. It has to be emphasized that this kind of analysis only indicates whether a stretch of the genome could be functional and not if it is indeed transcribed and thus a gene in the common sense. Computational classification of a putative ncRNA as a small trans-encoded antisense RNA, in the above described sense, is not possible with current knowledge.

SIPHT The sRNA identification protocol using high-throughput technology (SIPHT) uses a wide spectrum of different ncRNA gene characteristics to annotated new sRNA in unknown genome regions. The two most informative sources are the sequence conservation between different intergenic regions of different bacterial species and the signature of rho-independent termination hairpin (see page 39). Furthermore, potential associations with one of several transcription factor binding sites and homology to previously identified sRNAs are taken into account. Ideally, this leads to a specific annotation of new sRNA genes [128, 129]. Nevertheless, it also holds the disadvantage of missing ncRNA lacking a rho-independent terminator [168].

NAPP Nucleic acid phylogenetic profiling (NAPP) works with the basic assumptions that the sequence of ncRNA are conserved between different species and that ncRNA usually cluster on the genome. The first assumption seems unproblematic, whereas the second one, even if it holds for the current known ncRNA, have the strong potential to induce a bias in the ncRNA annotation.

To detect these clusters of conserved ncRNA, NAPP first defines conserved regions by applying a blast search for intergenic sequence tiles[6]. Conserved regions are checked against the Rfam database with all its annotated ncRNA [31]. If more conserved regions correspond to annotated ncRNA than expected by chance, the remaining conserved regions are also considered to be potential ncRNA [168].

RNAz In contrast to the two afore mentioned ncRNA detection approaches, RNAz [242, 77] regards the conservation and the stability of the structure and not the sequence of putative new ncRNA. The main advantage is the avoidance of any training set such as SIPHT, which can introduce the bias to annotate only more of the same already known sRNA types. The main disadvantage is that not all ncRNA are expected to come along with a conserved secondary structure. Especially antisense sRNA seem to function mainly via their sequence, although the accessibility and hence the structure is also important. This could be illustrated in [170], where it was shown that a 16 nt long sequence from the 5' end of *Salmonella* RybB attached to an unrelated ncRNA, which disrupts the original structure, remains functional and maintains its target specificity.

---

[6] Each intergenic region of a genome is cut into 50 nt pieces and blasted against all other reference genomes.

5.3 Functional characterization of small RNA

5.3.1 Experimental discovery of sRNA targets

Due to technological progress systematic screening of whole genomes for putative ncRNA genes became possible. In contrast, characterizing the functional role of an annotated sRNA in the gene regulatory network is still much more labor intensive and less suited for high-throughput screenings.

To gain a broad idea of its function the sRNA is often over-expressed or deleted and the resulting effect on the transcriptome is evaluated. In some cases the effects are strong enough that very simple read out methods are sufficient. The constitutive expression of the sRNA GlmY for example causes such an strong activation and subsequently an enrichment of its target GlmS protein that it was possible to detect it with a simple Coomassie-stained SDS gel [230].

A similar technique with more refined read out methods, was already successfully applied to screen whole genomes for sRNA involved in certain biological processes [103]. There, Jin et al. examined the role of sRNA in *E. coli*'s acid stress resistance. 79 different sRNA were deleted to construct a single-sRNA gene knockout library. This library was systematical tested for an altered acid resistance. Soon, the small RNA GcvB was identified to be involved in the regulation of the response to acid stress. It still took very laborious tests to distinguish between direct targets of gcvB and indirect effects on other genes. The testing of several double mutants eventually showed that all effects of gcvB deletion on the acid resistance disappear if a ΔgcvBΔrpoS mutant is compared with an ΔrpoS mutant. Thereby, rpoS was recognized as being directly activated by GcvB. All other effects on the transcriptome were explained by indirect effects subsequently caused by the altered rpoS activity [103].

Distinguishing between direct targets and indirect side effects is a general problem of experimental approaches with constitutive over-expression or deletion of sRNA genes. To overcome this issue sRNA genes can be recombinated and cloned under the control of an inducible promoter. This way it becomes possible to test the effect of an sRNA over-expression shortly after the onset of sRNA expression, reducing the time for secondary effects to propagate through the gene regulatory network [79].

Although this strategy proved to limit site effects and thus reduces the search space to identify direct targets, it was shown that already after very short time first indirect effect can become manifest in transcriptome profiles. For example, in one study [141] the transcriptome changes upon expression of the small regulatory RNA RyhB was evaluated 15 minutes after induction of the sRNA gene. Still, beside 56 putative direct targets[7] 29 genes[8] showed also an abundance

---

[7] Which are organized in 18 distinct operons.

[8] Organized in 10 operons.

change which are supposedly indirectly controlled [141].

5.3.2 Computational discovery of sRNA targets

To describe genes which are potentially influenced in their expression by one specific sRNA, generally this means describing mRNA which can physically interact with the small RNA by base pairing. The most basic in silico approach to detect them is searching for genes with a stretch of an imperfect complementary RNA sequence. Usually the length of interacting sequences is quite small, typically 9 nt up to 60 nt of imperfect complementarity is enough [160]. In the case of sRNA SgrS interacting with the mRNA ptsG it was shown that only 6 nt are critical for functional interaction [111]. This leads to a very high false positive rate. Hence, the search space must be confined somehow. Section 6.5 described the historic development of different approaches to make this in-silico target search more accurate. Chapter 8 and 9 elaborate new contributions conducted in the course of this doctoral thesis to the research field of computational sRNA target prediction.

5.3.3 Confirmation of sRNA targets

Once an idea of potential sRNA targets is established, either by experimental or computational methods, detailed characterization of the exact mechanism of action is desirable, although in general not achieved. This is due to the very labor intensive techniques to gain confidence.

To test a physical interaction in vitro and in vivo approaches can be applied or, ideally combined. The most common in vitro strategy is the determination of the sRNA–mRNA binding site by enzymatic or lead(II)-induced footprinting [40]. A technique, which is more suitable for automated high-throughput screenings, are in vitro translation systems, such as the minimal, cell-free PureSystem® (Cosmo Bio Co.) [51].

To confirm that a predicted interaction is indeed direct the technique of two-plasmid reporter gene assay has been established [231]. Thereby, the 5' UTR of a putative target and the first couple of codons are cloned in frame of a reporter gene, usually GFP. The sRNA is expressed from a second plasmid. With the reporter gene an easy read out of the target translation activity with and without expressed sRNA becomes possible.

To identify the exact bases which are involved in sRNA–mRNA hybridization, compensatory base-pair exchanges is considered the most reliable technique. There, bases in the sRNA are mutated until the regulatory function is lost. If the function is rescued by introducing compensatory mutations in the mRNA the exact binding site can be pinpointed [175].

Since all the experimental procedures are quite labor intensive, it is in the best interest to reduce the list of potential targets as far as possible. For this purpose experimental screening and computational analysis can come into consideration. The best strategy so far might still be a combination of as many techniques as possible, to eventually gain a complete picture of individual targets and a reliable understanding of the sRNA regulation system.

6 Computational RNA Biology

As most sciences, biology was revolutionized by the rapid development of new technologies over the past few decades. Maybe the most important was the introduction and vast application of computers. Therefore, two closely related but distinct disciplines developed, *Bioinformatics* and *computational Biology* (see definitions in Tab. 6.1). Computational Biology in this sense is more related to theoretical Biology. It aims to acquire insights into biological systems by simulating their behavior. Computational RNA Biology in particular applies this strategy to the function of RNA molecules in biology. It proved to be an especially fruitful field. This is due to the fact that the behavior of RNA can be characterized in many cases by its sequence and its structure. Both of which can be handled quite successfully with sophisticated techniques developed since the late 1970's [187].

Table 6.1: Definition of computational Biology and Bioinformatics according to [98].

Bioinformatics	Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
Computational Biology	The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

The current chapter aims to provide an introduction into the most pivotal challenge of computational RNA Biology, the accurate characterization of RNA secondary structure. Structure prediction is critical in almost all aspects when RNA is involved. This opening will be followed by an introduction into a selection of questions and applications, which serve as a base for chapter 8 and 9

6.1 In silico RNA folding

The prediction of secondary structure, as described in section 3.2, is a computational challenging task. Any method, which tries to tackle it, needs four essential elements. First, a general model of the *Architecture* of the RNA. There, it must be defined what features, such as stacked bases, canonical and non-canonical pairs, or loop length, contribute to the secondary structure in the model. Structures build on the basis of the architecture must be evaluated with an appropriate *Scoring scheme*. The score can be deduced from thermodynamic or probabilistic considerations, and must be specified with a defined set of *Parameters*. Architecture, scoring scheme and parametrization constitute the model of the secondary structure. Finally, the model must be applied onto a particular RNA sequence. This is executed by the *folding Algorithm*, to emit a description of plausible RNA secondary structure [187].

In the following subsections, I will first summarize how an RNA secondary structure can be conclusively described. Subsequently, the major innovations in the field of RNA secondary structure prediction will be explicated by means of the historical development leading to the recent, state-of-the-art structure prediction methods.

6.1.1 Secondary structure representations

Optimal structure Optimality is an ambiguous, hence controversially discussed, term, not only in computational RNA biology. In the field of in silico RNA folding, what is "optimal" depends on the used architecture and scoring scheme applied. Different attempts were already used, e.g. the best structure is the one with the maximal number of compatible base pairs [163]. Meanwhile, optimality is mostly defined by thermodynamic or probabilistic consideration. Applying a probabilistic scoring scheme, the most probable structure is the optimal structure [74]. Applying a thermodynamic approach, the minimal free energy structure can be considered to be the optimal structure.

It has to be emphasized that optimality can only hold in the limits of the applied model. If, for example, the architecture is not aware of pseudo-knots, the optimal structure will only be the best one without a pseudo-knot [214].

Boltzmann distribution and partition function A sample of RNA molecules can be described as a physical system, where every molecule can adopt a structure, or generally speaking, a certain state. On a larger scale and in the thermodynamics equilibrium the system can be described by the number of individual molecules residing in each state at a given moment. This is equivalent to the likelihood that a randomly chosen molecule will have a particular state adopted. Internal (e.g. sequence) and external (e.g. temperature) conditions influence this likelihood. It can be mathematically described with the Boltzmann distribution.

$$P(S_i) = \frac{N_i}{N} = \frac{e^{\frac{-E_i}{RT}}}{Z(T)} \tag{6.1}$$

In words, the probability P to find structure S_i equals to the proportion of N_i molecules with structure i within the complete set of $N = \sum_i N_i$ molecules, which in turn depends on the free energy E_i of structure i, the temperature T, the gas constant R, and the partition function $Z(T)$.

The partition function is a state variable of the RNA. In this sense, it tells something not just about one structure but about the whole ensemble of all possible structures[1]. In Eq. 6.1 it serves as an important normalization factor, granting that the sum of the probabilities over all structures $\sum_i P(S_i) = 1$, since $Z(T) = \sum_{S_i} e^{\frac{-E_i}{RT}}$. It has to be emphasized that although accurate, it is impractical to use this definition as an instruction to calculate the partition function. How this is practically achieved will be discussed in section 6.2.3.

Finding the partition function is related to enumerating all possible structures and can be interpreted as a weighted structure count. In the context of a thermodynamic parametrization the weight for each structure is its Boltzmann factor. In the case of a probabilistic implementation the weight reflects the probability of each structure [187].

The partition function becomes very useful, if the attention is shifted from the optimal structure to competing alternative structures. It enables not just to calculate the probability of one structure (e.g. the optimal structure) but also to express the probability that one specific base or a sub-sequence is engaged in base pairing [17]. From there, the probability for being unpaired, hence accessible to RNases or other RNA molecules, can be directly deduced.

Ensemble and maximal expected accuracy structure By using the partition function, it becomes possible to calculate the probability $P_{i,j}$ of any two bases i and j to pair with each other. This can be used for different purposes. For one, it enables to characterize the whole ensemble of potential relevant structures in a very clear way. The left hand side of Fig. 6.4 depicts a so called dot-plot, which illustrates all possible inter-molecular base pairs with their corresponding probability. It might seem, on the first sight, harder to interpret than the wide spread structure drawings (Fig. 6.4, r.h.s), but it holds so much more information and has the potential to change "*the heavy but misguided emphasis on single unique structures for biological macromolecules*"[2] [144].

Furthermore, it can be used to define the maximal expected accuracy (MEA) structure as the structure with the highest sum over all single base pairs $\sum P_{i,j}$. The MEA structure was shown to perform slightly better in reproducing experimentally determined structures than the corresponding MFE structure [80].

---

[1] Once again, "possible" with the restriction of the applied model.

[2] This are promising words, verbalized over twenty years ago by McCaskill [144]. To this day the change he hoped for, has not yet arrived. May this recurrence serve to this end.

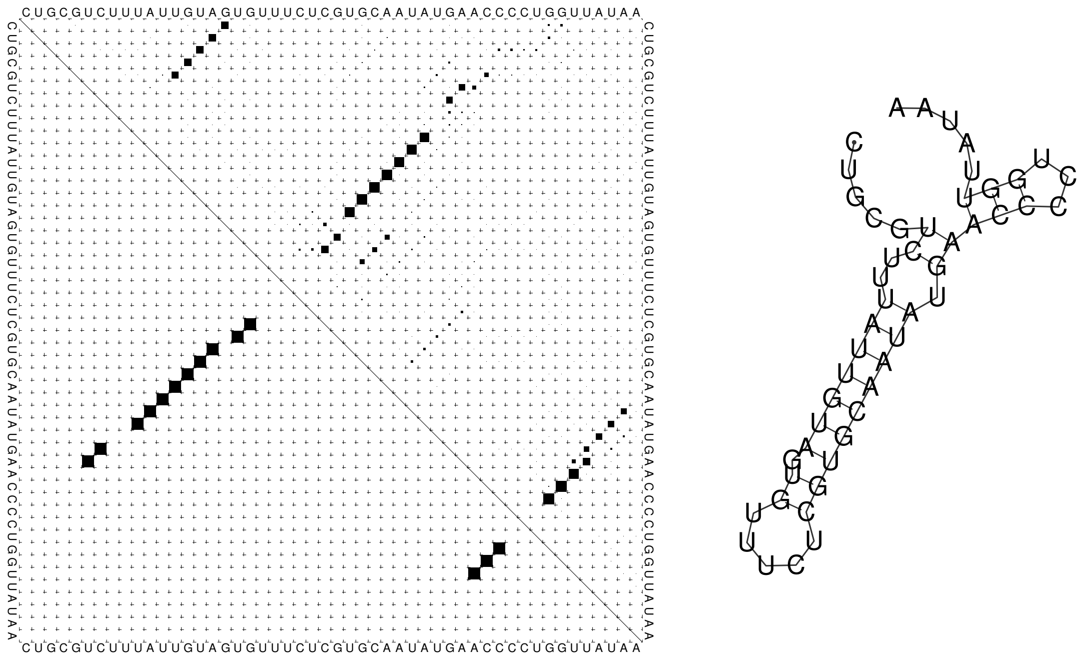


Figure 6.1: Dot-plot (l.h.s.) representing the structure ensemble of a 50 nt random sequence. The axis are labeled with the bases. The lower triangle represents the MFE structure (also illustrated on the r.h.s.). Each filled square indicates that the corresponding base in the x-axis pairs with the counterpart from the y-axis. The upper triangle represents the more informative ensemble of structures. Here, the size of the filled square represents the relative frequency of this particular base pair in the ensemble of all structure. It reveals that beside the two MFE stems (the longer with an interior loop and a bulge), there are other unstable structures which could have biological relevance. Example was calculated and plotted by `RNAfold -p`.

6.2 RNA folding algorithm

6.2.1 Nussinov and Jacobson

Algorithm-wise the most important development leading to nowadays RNA folding programs was already achieved around 35 years ago, when Ruth Nussinov et al. introduced their algorithm to efficiently calculate the structure with maximal matching size [163]. Therefore, they restricted the architecture of the structure to a planar loop graph and applied a dynamic programming algorithm to find the optimal graph.

Planar loop graph Nussinov et al. visualized the problem of finding the structure with the maximal number of base pairs for a given sequence by a loop graph with certain restrictions facilitating the calculation of the solution. In this graph the vertices, representing the single nucleotides, are arranged in a circle connected by the edges representing the ribose phosphodiester backbone. Each vertex can form an additional edge, representing base pairing interactions. For these edges the following constraints have to be fulfilled [163]:

- Two vertices can only be connected if the two base species respect the pairing rules, i.e. A-U, C-G, G-U, U-A, G-C or U-G.
- A vertex must not be engaged in more than one base pairing edge. This restriction excludes many non-canonical interaction such as G-quadruplexes [133].
- Two neighboring vertices in the backbone must not pair. This is in agreement with steric consideration of the flexibility of the RNA backbone.
- The base pairing edges can only be drawn in the interior of the loop and must not cross each other, that is why the loop graph is called planar. This restriction exclusively allows for nested structures and excludes pseudo-knots [177].

Dynamic programming Many optimization problems can be solved by recursively solving sub-problems. In the course of this strategy one frequently faces the same sub-problem over and over again. Re-solving the same sub-problems is computational very inefficient. Dynamic programming can circumvent this by tabulating the solutions for the sub-problems. In case of reappearing, the solution can be looked up instead of being recalculated, which generally leads to a huge improvement in the run time performance at the expense of higher memory demand. In the case of RNA secondary structure prediction it leads to a reduction of time complexity, since the number of possible structures S_n for an RNA sequence with length n, which have to be evaluated for optimality, grows exponentially with $S_n \propto n^{(-\frac{3}{2})}\alpha^n$ with $\alpha \simeq 1.85$ [92]. Whereas, the number of steps filling the recursion matrix growth only with the power of 3, leading to a time complexity for the algorithm of $\mathcal{O}(n^3)$. The matrix size itself needs $\mathcal{O}(n^2)$ memory space[3].

In the particular case of recursive RNA folding, the problem of finding the structure with the maximal number of base pairs for a sequence with length n, can be deconstructed into the sub-problem of finding optimal solutions for sub-sequences. In the course of the recursive calculation a matrix $\mathcal{M}_{[1,n]}$ is filled step-wised, whereas the entry $\mathcal{M}_{[i,j]}$ represents the optimal score for the sub-sequence $s_{[i,j]}$ (Eq. 6.2). Given $\mathcal{M}_{[i-1,j]}$ the optimal structure for $s_{[i,j+1]}$ can be deduced by finding the maximum of the possible extensions, which are (i) the added vertex is not involved in a base pair, (ii) the add vertex prolongs the structure from $\mathcal{M}_{[i,j-1]}$ by one base pair, or (iii) the optimal structure of the sub-sequence $s_{[i,j]}$ is the additive combination of

---

[3] The authors of [163] commented this, considering the computational power at their disposal, with: "While the time requirement of $\mathcal{O}(n^3)$ (on the order of n^3 elementary operations) of the resulting algorithm is reasonable even for values of n of a few thousand, the $\mathcal{O}(n^2)$ memory requirements are somewhat of a practical difficulty."

two compatible substructure $\mathcal{M}_{[i,k]} + \mathcal{M}_{[k+1,j]}$, whereas x_i and x_j can form a base pair, thus are element of the set of allowed base pairs $\mathcal{B} = \{AU, UA, GU, UG, GC, CG\}$.

$$\mathcal{M}_{[i,j]} = \max \begin{cases} \mathcal{M}_{[i+1,j]} \\ \max\limits_{k}\{\mathcal{M}_{[i+1,k-1]} + \mathcal{M}_{[k+1,j+1]} + 1\} & \text{with } i < k < j \text{, and } (x_i, x_j) \in \mathcal{B} \end{cases} \tag{6.2}$$

Once the matrix $\mathcal{M}$ is filled, the score of the optimal structure is known. In a final step the process can be reverted. From the tabulated values of all sub-solutions the path to the optimal score, and hence the optimal structure can be reconstructed. This process is called back tracking.

The scoring scheme applied in the basic Nussinov algorithm is based on simple weights for each base pair, whereas each allowed and formed base pair has the same weight (which is the simplest parametrization possible) [163].

6.2.2 Zuker and Stiegler

Although the algorithm by Nussinov et al. excelled in its sophisticated and efficient way to calculate the "optimal" structure, it became clear that the assumption, optimality can be scored by simply counting the base pairs, does not produce reliable predicted structures [187]. That is why, soon after Zuker and Stiegler seized the algorithmic aspects of the approach and worked out a considerable refinement of the scoring scheme and parametrization [257].

The comprehension that RNA folding is guided by thermodynamic processes, stood in the beginning of the new scoring scheme. Thereby, an energetically stable structure is more likely to be formed. The use of thermodynamic scoring scheme and parameters were already suggested and applied before, e.g. in [176, 217]. The achievement of Zuker et al. was its incorporation into an efficient algorithmic framework by fusing it with the dynamic programming algorithm by Nussinov.

Therefore, the property that the energy of an RNA structure can be expressed as the sum of the energies of its substructures was used. To account, beside the base pairing energies, also for the stacking energies, the fundamental unit at the end of the decomposition is not the base pair but different loops. The energy for the most common loop types, such as hairpin, interior and exterior loops (see Fig. 3.4 on page 15) can be measured for representative structures and extrapolated for the remaining [63]. Multiloop features are unfortunately not accessible for measurements and have to be frankly guessed [187]. The overall free energy $E(S)$ of a structure S decomposed this way can be calculated by summing up the energy contributions $E(L)$ overall loops L structure S consists of.

$$E(S) = \sum_{L \in S} E(L) \tag{6.3}$$

Applying this new scoring scheme together with a dynamic programming algorithm is in principle similar to the recursions described in subsection 6.2.1. Although, to comply with the more complex scoring scheme it becomes a bit more elaborated, since additional recursion matrices must be considered. The recursion can be written as in Eq. 6.4 (adapted from [257]).

$$
\begin{aligned}
\mathcal{V}_{[i,j]} =& \min \begin{cases} \mathcal{H}_{[i,j]} & \\ \min\{\mathcal{J}_{[i,j;i',j']} + \mathcal{V}_{[i',j']}\} & \text{with } i < i' < j' < j \\ \min\{\mathcal{W}_{[i+1,i']} + \mathcal{W}_{[i'+1,j-1]}\} & \text{with } i+1 < i' < j-2 \end{cases} \\
\mathcal{W}_{[i,j]} =& \min \begin{cases} \mathcal{W}_{[i+1,j]} & \\ \mathcal{W}_{[i,j-1]} & \\ \mathcal{V}_{[i,j]} & \\ \min\{\mathcal{W}_{[i,i']} + \mathcal{W}_{[i'+1,j]}\} & \text{with } i < i' < j-1 \end{cases}
\end{aligned}
\tag{6.4}
$$

Here, the matrix $\mathcal{W}_{[i,j]}$ stores the free energy for the optimal substructure of sub-sequence $s_{[i,j]}$. Similarly, $\mathcal{V}_{[i,j]}$ stores the free energy for the optimal substructure between i and j with the constraint that i and j must pair with each other. The table $\mathcal{H}_{[i,j]}$ holds tabulated free energy of a hairpin loop with closing pair (i, j). $\mathcal{J}_{[i,j;i',j']}$ is the tabulated free energy of an internal loop with the unpaired sub-sequences $s_{[i,i']}$ and $s_{[j',j]}$. The last option considered for minimization corresponds to a decomposition of a multiloop into two independent optimal substructures $\mathcal{W}_{[i+1,i']}$ and $\mathcal{W}_{[i'+1,j-1]}$.

The recursive computation of the matrix $\mathcal{W}_{[i,j]}$, storing optimal sub-solutions for the subsequences $s_{[i,j]}$ without the constraint that (i, j) must pair, is done in a similar way. This time existing optimal structures are, if energetically favorable, extended by an unpaired base.

The decomposition in the above form is not non-ambiguous since it does not take care how exactly multiloops are split into optimal substructures if they have the same overall score. For the calculation of the optimal structure this does no harm, for the calculation of the partition function it has to be avoided since it can not guarantee each structure is considered, and is considered only once.

The recursion in Eq. 6.4 can also be interpreted graphically as the optimal decomposition of the sub-sequence $s_{[i,j]}$ into different loops from the set of considered loop types. In this sense, the minimization options in Eq. 6.4, line 1 to 7 can be interpreted such that the one option gets chosen which represents the optimal decomposition. For each option its decomposition interpretation is given below.

$\mathcal{H}_{[i,j]}$ The optimal structure between i and j, with the base pair (i, j), is a hairpin loop, i.e. an U-turn loop closed by base pair(s) (Fig. 3.4, first from the left).

$\mathcal{J}_{[i,j;i',j']} + \mathcal{V}_{[i',j']}$ The optimal structure between i and j, with the base pair (i,j), is an interior loop or a bulge, i.e. unpaired bases between two closing base pair(s). Possibly asymmetric (Fig. 3.4, second from the left).

$\mathcal{W}_{[i+1,i']} + \mathcal{W}_{[i'+1,j-1]}$ The optimal structure between i and j, with the base pair (i,j), is a multiloop consisting of two optimal sub-substructure of any type (Fig. 3.4, first from the right).

$\mathcal{W}_{[i+1,j]}$ The optimal structure between i and j is the optimal structure of $s_{[i+1,j]}$ extended by one unpaired base.

$\mathcal{W}_{[i,j-1]}$ The optimal structure between i and j is the optimal structure of $s_{[i,j-1]}$ extended by one unpaired base.

$\mathcal{V}_{[i,j]}$ The optimal structure between i and j is equal to the optimal structure of $s_{[i+1,j]}$ of any type, providing that i and j pair with each other.

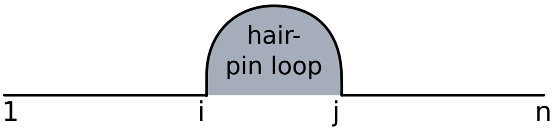
$\mathcal{W}_{[i,i']} + \mathcal{W}_{[i'+1,j]}$ The optimal structure between i and j is an exterior loop (Fig. 3.4, second from the right)

Computational complexity Compared to the less elaborated Nussinov algorithm the more complex scoring scheme of the Zuker algorithm is reflected by a higher computational complexity. The time complexity for a trivial implementation, without optimization and applied heuristics, grows from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^4)$. This increase is caused by the evaluation of possibly optimal interior loops, because there neither stem length nor bulge size is restricted. The memory complexity is still $\mathcal{O}(n^2)$, but due to a second upper triangular matrix it require at least twice as much compared to the Nussinov approach.

6.2.3 McCaskill

Similar to the already presented approaches, revealing new structure properties from an RNA sequence is achieved in many cases by dynamic programming. This is also true for the first applicable implementation of an algorithm which efficiently calculates the partition function, which in turn can be used to compute equilibrium probabilities of structures and base pairs. McCaskill proposed and applied such an approach in 1990 [144]. In contrast to the Zuker algorithm which is an optimization algorithm, McCaskill's goal is the optimal computation of the huge number of different possible structures, given a certain sequence.

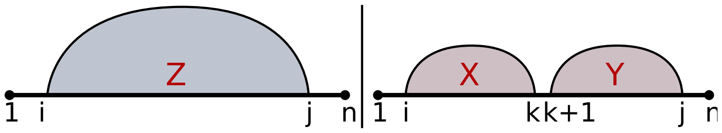
The main difference to the Zuker MFE algorithm is the different nature of the parameters in use. There, the minimal energies of the sub-solutions, which correspond to an optimization procedure and are additive to gain the energy of the total structure. In contrast, the sub-partition functions are multiplicative for disjoint structure [144]. Two structures are disjoint if they do not share a base pair. For example, if the partition function including a base pair between i and j which closes a hairpin loop is calculated (depicted in Fig 6.2), one has to consider the sub-partition functions of all segments, and multiply each term.



$$Z_{1,n} = Z_{1,i-1} \cdot Z_{i+1,j-1} \cdot Z_{j+1,n} \cdot e^{\frac{-E(i,j)}{RT}}$$

Figure 6.2: Reconstructing the partition function contributions for a simple hairpin loop. In contrast to the energy calculation where the sub-terms would add up to the total energy, the sub-terms of the partition function multiply each other. In this simplified form the equation is only valid for the Nussinov scoring scheme, since it is applied on a base pair instead of a loop.

If two alternative structures are independent from each other, e.g. multi loop or a hairpin on the same sub-sequence, an MFE algorithm chooses the substructure with the smallest energy contribution, in other words the optimal substructure is chosen. To calculate the partition function, the sub-partition functions of the possible substructures are summed up.



$$Z_{1,n} = (Z_{1,i-1} \cdot Z_{i,j} \cdot Z_{j+1,n}) + (Z_{1,i-1} \cdot Z_{i,k} \cdot Z_{k+1,j} \cdot Z_{j+1,n})$$

Figure 6.3: For alternative structures on the same sub-sequence the partition function of each structure contributes additive to the complete partition function. In the depicted example between i and j either one embracing hairpin loop or two individual hairpin loops can be incorporated. Partition functions of disjoint substructures, such as X to Y, are multiplied. Independent, alternative structures' partition functions, like Z to $X + Y$, are summed up. This is only valid if the decomposition is done in a non-ambiguous way, like proposed by McCaskill and implemented in the `RNAfold` program.

With these two modifications the dynamic programming approach by Zuker can be adapted to calculate the partition function. Again, this leads to an overall run-time complexity of $\mathcal{O}(n^4)$. A more sophisticated implementation achieves a complexity reduction by restricting the interior loop sizes to some constant value. This way the partition function, and from there the probability of a particular structure or base pair, can be calculated with a run-time complexity of $\mathcal{O}(n^3)$.

RNAfold Many modern RNA folding programs are in its core still derivations of the algorithm and the scoring scheme introduced by Zuker et al. and McCaskill. This also includes the widely used program `RNAfold` [91] shipped with the `ViennaRNA package` [131].

6.3 Local RNA folding

All so far discussed RNA folding strategies have in common that they examine an input RNA sequence globally. For some application, however, a local structure prediction is more favorable. Partly, this is caused by technical issues. For instance, global folding of very long sequences is still not feasible with common computer resources. Furthermore, many ncRNA genes are not well defined in terms of their exact boundaries. Likewise, the exact genomic start and stop coordinates of the majority of bacterial cistrons are not yet determined. Since the outcome of global folding algorithms is critically dependent on the input sequence, already small uncertainties of the sequence boundaries might have great influence on the predicted structures.

Beside, there are also physical reasons to favor a local folding of an RNA sequence. On one hand, in the context of the cytosol, it is reasonable to assume that any RNA molecule is associated with RNA binding molecules. *E. coli* mRNA, for example, were shown to be covered by ribosomes, leaving on average only 46±6 nt from the exit site of the mRNA of one ribosome to the entrance site of the next ribosome [26]. Only this stretch can form a secondary structure. Base pairs enclosing a whole ribosome are conceivable, but in this case the applied energy parameters are certainly not suited to account for that. On the other side, it was also shown that kinetic aspects of RNA folding discriminate against long range interactions in favor of short range base pairs [62]. This might be one reason why global thermodynamics-based RNA secondary structure predictions still suffer from a lower accuracy in long-range base pair prediction compared to the accuracy of short-range interactions [164, 57].

6.3.1 RNAplfold

Some of the aforementioned limitations can be overcome if a local folding strategy is applied. One such an implementation is the program `RNAplfold` [17] from the `ViennaRNA package` [131]. The program does not emit one predicted structure, but instead the local base pair probabilities for base pairs with a maximal span of L within a sliding window of size W. Therefor the pair probability is calculated in a window of size $W \geq L$. Eventually the probability of base pair (i, j) is averaged over all windows of size L which contain this base pair. Generally speaking, `RNAplfold` applies a recursion scheme to calculate the average equilibrium probability of a base pair (i, j) over all fixed-size sequence windows [17].

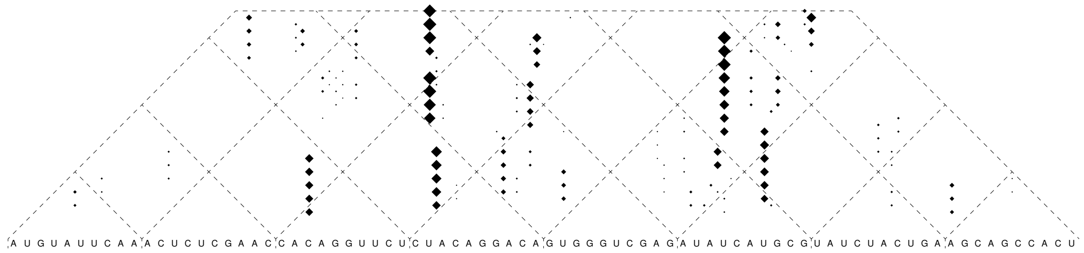


Figure 6.4: Example of an `RNAplfold` output for a 80 nt long random sequence calculated with the parameters `-W 35 -L 34`. The output consists of a dot plot in postscript format. The area of each black square represents the averaged pair probability of the two corresponding bases. Stable local stems appear as vertical columns.

The output consists of a visualization of the averaged base pair probabilities in form of a dot plot, which can easily be parsed and visually inspected (Fig 6.4). Further more, the averaged probabilities that a stretch of n consecutive nucleotides is unpaired are provided. This feature becomes useful for the prediction of binding sites (see chapter 8).

6.4 Consensus structure for RNA alignments

Despite the advances in thermodynamics-based RNA secondary structure prediction the accuracy of the computed structures is somehow limited. This can be explained by several factors. For one, efficient computation of the RNA fold relies on the restriction to nested structures, omitting e.g. pseudo-knots. Even if in thermodynamic equilibrium the MFE structure is the most likely and hence the most occurring structure, biologically interesting structures are sometimes metastable, far away from the equilibrium state. Kinetic considerations might be more important, especially if the biological function is associated with structure changes in the course of action, e.g. riboswitches. And finally, despite the prudence and the volume of measured thermodynamic parameters used in the models, they still have to be regarded as sophisticated estimations, introducing some uncertainty [134]. This leads to an overall prediction accuracy of 50 to 70 % [52, 54].

Different strategies to overcome these limitations of RNA sequence folding were proposed. Beside tweaking the parameters [6] and using more general and abstract structure representations, such as centroid [55] or MEA structure [80], the incorporation of additional information into the in silico folding process showed to be promising approaches.

For one, additional data from in vitro experiments can be used. Such an experiment typically consists of probing the RNA of different RNases with different specificity for certain motifs, such as hairpin or free loops, and certain bases. These experiments do not reveal the complete picture for the structure but the produced information can be used to guide the folding procedure into the right direction. One way to do this is by introducing so called soft constraints to favor certain bases with a pseudo energy if they are in agreement with the experimental data [243].

An other successfully used source of information which is in some cases easier to obtain than experimental data, is the consideration of conserved structures. If the sequence of interest is conserved in other, related species, this conservation information can be utilized by providing an alignment instead of a single sequence as input for the folding algorithm. Beside others, e.g [205, 48], one implementation of such an approach is `RNAalifold` [93, 19] part of the `ViennaRNA package` [131].

6.4.1 RNAalifold

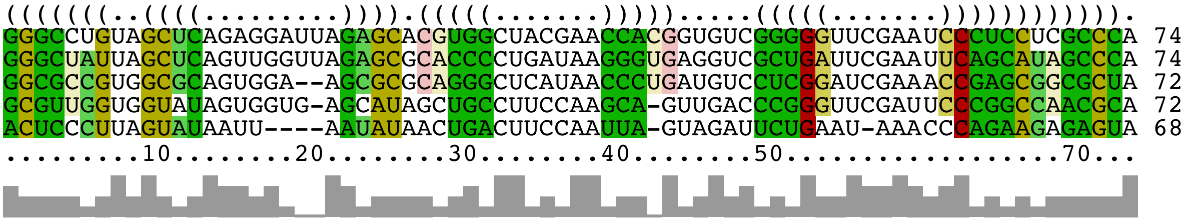
`RNAalifold` calculates the consensus structure for a provided RNA alignment considering the base mutation pattern in the provided sequences. Under the assumption that the function of an RNA is mediated by its structure, the structure can be expected to be conserved. Similar to protein coding genes with their neutral mutations[4], in ncRNA genes a base substitution can be silent if it does not break the resulting functional structure. In this context "consistent mutation" and "compensatory mutation" can be distinguished. The first corresponds to a single mutation which preserves the base pairing, e.g. if the C in the base pair G–C mutates to a U the bases G–U can still pair. The latter corresponds to a mutation and a second compensatory mutation, which restores the base pair. For example, if the base pair G–C changes to the base pair A–U via two single mutation (G→A and C→U), the pairing is not broken. Both mutation patterns indicate that the particular base pair is functional important because there seems to be a selection pressure conserving it.

In its core, `RNAalifold` applies the same folding algorithm as `RNAfold` on the alignment columns instead of the sequence positions, with some important differences. For one, the rule that only canonical base pairs can be formed is changed to account for different bases in an alignment column. Pairs between column i and j can be formed if a predefined fraction of bases in i can interact with the corresponding base in j. Second, each closed base pair is evaluated for its energy and its covariance contribution simultaneously in the recursive calculation of the optimal structure. The covariance is either measured as the sum of all hamming distances between the

---

[4] The substitution of a nucleotide and its corresponding codon triplet to different triplet which encodes for the same amino acid according to the genetic code is called neutral or silent mutation, since it does not change the gene product, the encoded protein.

bases in each column [93] or based on a RIBOSUM[5] measure [19]. The first strategy gives a bonus for each mutation with preserved base pairing, favoring complementary over consistent over no mutation. The RIBOSUM matrix approach is quantitatively more sound. It appoints for each base pair a log-odds score how likely this conversion is. These probabilities were trained on a data set of 13,500 ribosomal sequences clustered into sets of similar evolutionary distance [19]. The last difference introduced by an input alignment, instead of a single sequence input, is how gaps are treated. In the first `RNAalifold` version [93] gaps were not treated differently than bases. In the later version [19] gaps are removed for each sequence individually before calculating the energy contribution of loops. Using the real loop sizes for the sequence instead of the virtual loop size in the alignment enables to retrieve the correct energy parameters for each sequence.



```
(((((((..((((.........)))).(((((.......))))).....(((((.......))))))))))))).
GGGCCUGUAGCUCAGAGGAUUAGAGCACGUGGCUACGAACCACGGUGUCGGGGGUUCGAAUCCCUCCUCGCCCA 74
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA 74
GGCGCCGUGGCGCAGUGGA--AGCGCGCAGGGCUCAUAACCCUGAUGUCCUCGGAUCGAAACCGAGCGGCGCUA 72
GCGUUGGUGGUAUAGUGGUG-AGCAUAGCUGCCUUCCAAGCA-GUUGACCCGGGUUCGAUUCCCGGCCAACGCA 72
ACUCCCUUAGUAUAAUU----AAUAUAACUGACUUCCAAUUA-GUAGAUUCUGAAU-AAACCCAGAAGAGAGUA 68
.........10........20........30........40........50..................70...
```

Figure 6.5: Consensus structure for alignment-folding calculated by `RNAalifold` for a showcase sequence alignment, derived from the sub-sequence of plant tRNA^{Arg}. The color tone indicate the number of different base pair types (increasing from red over ocher to green). The color hue gives the number of incompatible pairs. The program was called with parameters `-p -color -aln`.

Eventually, `RNAalifold` emits a consensus structure and the underlying base pair probabilities for the provided input alignment, whereas the total energy of the alignment-folding consists of the sum of averaged loop-based energies of all sequences and the covariance contribution [19].

Application The implementation of `RNAalifold` described above comes with a time complexity of $\mathcal{O}(Nn^3)$, whereas N are the number of sequences and n number columns in the alignment. Hence, the algorithm is efficient enough to screen very long sequences for potentially conserved structures. The most natural scope is its application to detect weak but well conserved signals. Chapter 8 will, among others, examine the application of a related methodology to identify conserved interaction sites between bacterial sRNA and their mRNA targets.

---

[5] RIBOSUM: ribosomal rna substitution matrix [114]. Inspired by its protein equivalent BLOSUM: blocks substitution matrix [89]

6.5 sRNA target prediction

As outlined in chapter 5 small RNA are an important participant in the coordinated cellular response to external stimuli. The number of newly discovered sRNA genes is constantly rising, making their functional characterization an immediate need. Since the mode of action is predominantly mediated by binding their targets via base pairing, functional characterization can be seen, from the biocomputational point of view, as reliable determination of the secondary hybrid structure of the two RNA molecules. RNA–RNA interactions follow the same mechanisms, such as base pairing and stacking, which also guide single sequence folding. Although physically equivalent, it is useful to distinguish between intra-molecular base pairs, i.e. within one sequence, and inter-molecular base pairs, i.e. between two sequences, since they are treated differently by some programs.

Due to these similarities the tool set of in silico RNA folding can be applied to detect new sRNA binding sites and the corresponding targets. Unfortunately, there are some obstacles which complicate the matter. One main difference in the physical properties is concentration dependence. For intra-molecular interactions each pair of potentially interacting sub-sequences can be found in a one to one concentration ratio in the system. For bi-molecular systems this assumption might not reflect realistic biologic conditions [53].

Technical differences are especially hindering. The scope of the developed folding algorithms are mostly to fold sequences up to several thousands nucleotides. To detect targets for an sRNA it is desirable to screen millions of bases for potential local RNA–RNA interactions. This needs special adaptations to the computational strategies. A variety of different approaches were proposed to reduce the complexity of the task in such a way that the scanning of complete genomes becomes feasible. In the following, selected examples are presented focusing not so much on technical details but more on how the balance between efficiency (reduced model complexity) and realistic physics (increased model complexity) is handled.

6.5.1 Hybridization

One approach, mainly focusing on efficiency, is to concentrate solely on the hybridization aspects of the RNA–RNA interaction. The very crudest approach is to simply scan for reverse complementary regions between the target and the query sequence in a BLAST-like manner [3]. A more proper way is to score the potential of base pairing interaction based on a thermodynamic model, instead of scoring similarities. This implementation is inspired by the Smith-Waterman dynamic programming approach for local alignments [208]. It leads to a very time and memory efficient behavior. This approach was first implemented in `RNAhybrid` [184] intended to predict miRNA targets. There, the recursions are adapted to forbid intra-molecular base pairs, limiting the maximal loop size, and neglect multiloops. This leads to a run-time complexity of $\mathcal{O}(mnc^2)$, depending on the query length m, target length n and maximal loop size c. The `RNAhybrid` algorithm was recycled in `RNAduplex` and the web-service `targetRNA` [223], which, due to its usability, is widely used. Later on, `RNAplex` [219] and `RIsearch` [245] were developed in the same spirit with a considerably refinement of the model.

Hybridization focused approaches benefit from the very fast run time, making them suitable to screen whole transcriptomes for potential targets. Their major drawback however, is the neglect of intra-molecular structures. Every inter-molecular base pair has to compete with many possible intra-molecular base pairs. Omitting this kind of information often leads to an overestimation of the interaction length, and as a consequence the corresponding binding energy. This impedes a reliable ranking of putative binding sites in order of their reliability.

6.5.2 Co-folding by concatenation of two sequences

On the one hand, including intra-molecular base pairing inflates significantly the complexity of the computational task. On the other hand, the capability of a putative binding site to engage in an inter-molecular binding critically dependents on how tightly this region is already engaged in intra-molecular base pairing. Having this information available can shrink the search space considerably. The first attempts to use this information was done by the computationally straight forward strategy of concatenating the target and query sequence via an artificial linker and apply regular folding approaches to this virtual molecule. In the beginning the same recursions as in the single sequence folding were applied. Later on, the loop which contains the linker was treated differently (e.g. `RNAcofold` [18], `PairFold` [5]).

The resulting implementation has the same time complexity as single sequence folding, $\mathcal{O}((n+m)^3)$ with n and m being the length of the query and the target sequence. The major disadvantage, however, is that the restriction to nested structures for the concatenated sequence $n+m$ disallows so called kissing-hairpins, which are a kind of a virtual pseudo-knot. Virtual in the sense that each sequence is knot free but the fold of the virtually concatenated sequence includes a pseudo-knot. This type of structure, where two intra-molecular hairpins form an inter-molecular interaction with the unpaired loop region, is not uncommon for sRNA–mRNA binding.

6.5.3 Accessibility aware hybridization

Combining the two advantages of sequence concatenation and hybridization, namely the consideration of intra-molecular structures, and the consideration of virtual pseudo-knot like structures, can be achieved when the RNA hybridization is modeled as a two step process. First, the complete base pair probabilities of the query and the target sequence are calculated. From there the energy which is needed to unfold a binding site ΔE_{open} can be directly deduced. In the next step the energy gained by the hybridization of the two binding sites ΔE_{hybrid} is calculated. The overall binding reaction is thermodynamically driven by the cumulative net energy.

$$\Delta\Delta E = \Delta E_{open}^{target} + \Delta E_{open}^{query} + \Delta E_{hybrid} \tag{6.5}$$

The calculated $\Delta\Delta E$ net binding energies allow a ranking how efficient a putative binding site can be expected to be used.

This approach is implemented in `RNAup` [159] and `intaRNA` [32]. The model is closer to the physical reality of RNA–RNA interactions, but comes with the costs of longer run-time. `RNAup` shows a run-time complexity of $\mathcal{O}(n^3 + m^3 + n \cdot m \cdot w^4)$. Thereby, n, m are the sequence length of query and target, and w the length of the maximal considered binding site.

The interaction model implemented in `RNAup` and `intaRNA` results in the concentration of intermolecular base pairs in a region which is not interrupted by intra-molecular base pairs. This neglects, for example, double kissing hair pins, such as the OxyS–fhlA interaction [7]. Nevertheless, both interaction regions, in the target and on the query, can be nested in internal structures, allowing single kissing hair pins. This limitation is eliminated by a refined approach called `biRNA` [42], which calculates not only the partition function of the hybridization but includes the full interaction partition function. Once again, this leads to an increase in run-time to $\mathcal{O}(n^4 \cdot w + m^4 \cdot w + n^2 \cdot m^2 \cdot w^4)$. As above, m and n are the sequence lengths of query and target, and w the maximal binding site length [42].

In [42] the authors report that the calculation of one sRNA (sequence length between 71 and 253 nt) and the 500 nt wide region around the start codon of the mRNA takes between 10 minutes to slightly more than an hour. This makes `biRNA` hardly suitable to scan complete transcriptomes in an acceptably time. Similar, although less pronounced, is the situation for `RNAup`

6.5.4 Outlook

All above mentioned methods suffer from a high false positive prediction rate, mainly because of the simplified energy functions and restricted types of considered interaction models. So far, rougher, thus less accurate, models had to be applied to grant a run time of the algorithms which allowed to screen all genes from a genome for putative target mRNA. Since this trade-off is somehow dissatisfying, in chapter 8 a new contribution is presented tackling this problem. The challenge here was to provide the accuracy of elaborated energy models and the speed of simplified scan. Furthermore, chapter 9 introduces a new model to simulate the translation initiation process. It enables to examine prior proposed sRNA binding sites whether they have potentially an effect on translation initiation. Since this is the main mode of action of regulatory small RNA, applying the model as a post-analysis step to the RNA–RNA interaction prediction can further increase the specificity of sRNA target prediction.

Part III

Publication

7 Transcription start site annotation using dRNA-seq data

7.1 Statement of personal contribution

SF, PFS, ILH, FA designed the study. FA programmed the back-end. MTW, RL, FA programmed the web service. SF programmed the front-end. FA wrote the paper.

7.2 Article

Fabian Amman, Michael T. Wolfinger, Ronny Lorenz, Ivo L. Hofacker, Peter F. Stadler, and Sven Findeiß.
"TSSAR: TSS Annotation Regime for dRNA-seq data."
BMC Bioinformatics xx, no. xx (2013): xx - xx

Received on June 19, 2013,
Revised on November 29, 2013,
Accepted on xxx;

© 2010 by the authors; licensee BioMed Central; Reprint with permission from BioMed Central, and the authors.

TSSAR: TSS Annotation Regime for dRNA-seq data

Fabian Amman [1,2] and Michael T. Wolfinger [2,3,4] and Ronny Lorenz [2] and Ivo L. Hofacker [2,5,9] and Peter F. Stadler [1,2,5,6,7,8] and Sven Findeiß[2,9]

[1] Bioinformatics Group, Department of Computer Science and the Interdisciplinary Center for Bioinformatic, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany.
[2] Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria.
[3] Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Dr. Bohr-Gasse 9, 1030 Vienna, Austria.
[4] Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, University of Vienna, Dr. Bohr-Gasse 9, 1030 Vienna, Austria.
[5] Center for RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark.
[6] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.
[7] Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany.
[8] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501.
[9] Research group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria.

Email: Fabian Amman*- afabian@bioinf.uni-leipzig.de;

*Corresponding author

Abstract

Background: Differential RNA sequencing (dRNA-seq) is a high-throughput screening technique designed to examine the architecture of bacterial operons in general and the precise position of transcription start sites (TSS) in particular. Hitherto, dRNA-seq data were analyzed by visualizing the sequencing reads mapped to the reference genome and manually annotating reliable positions. This is very labor intensive and, due to the subjectivity, biased.

Results: Here, we present `TSSAR`, a tool for automated *de novo* TSS annotation from dRNA-seq data that respects the statistics of dRNA-seq libraries. `TSSAR` uses the premise that the number of sequencing reads starting at a certain genomic position within a transcriptional active region follows a Poisson distribution with a parameter that depends on the local strength of expression. The differences of two dRNA-seq library counts thus follow a Skellam distribution. This provides a statistical basis to identify significantly enriched primary transcripts.

We assessed the performance by analyzing a publicly available dRNA-seq data set using `TSSAR` and two simple approaches that utilize user-defined score cutoffs. We evaluated the power of reproducing the manual TSS annotation. Furthermore, the same data set was used to reproduce 74 experimentally validated TSS in *H. pylori* from reliable techniques such as RACE or primer extension. Both analyses showed that `TSSAR` outperforms the static cutoff-dependent approaches.

Conclusions: Having an automated and efficient tool for analyzing dRNA-seq data facilitates the use of the dRNA-seq technique and promotes its application to more sophisticated analysis. For instance, monitoring the plasticity and dynamics of the transcriptomal architecture triggered by different stimuli and growth conditions becomes possible.

The main asset of a novel tool for dRNA-seq analysis that reaches out to a broad user community is usability. As such, we provide `TSSAR` both as intuitive RESTful Web service (http://rna.tbi.univie.ac.at/TSSAR) together with a set of post-processing and analysis tools, as well as a stand-alone version for use in high-throughput dRNA-seq data analysis pipelines.

Keywords: differential RNA sequencing, dRNA-seq, TSS, Transcription start site annotation, Transcriptome, RESTful Web service, next generation sequencing

Background

Deep sequencing approaches were successfully applied to examine the architecture of primary bacterial transcriptomes and uncovered an unexpectedly complex achitecture [1–5]. Although plain transcriptome sequencing can in principle be sufficient to determine transcription start sites (TSS) as local accumulations of read starts, this approach requires extensive sequencing depth [6,7]. Alternative TSS located within well-expressed genes or operons remain undectable since moderate changes in coverage do not offer a sufficiently distinctive signal. On the other hand, TSS are not the only loci at which read starts accumulate in RNA-seq data. Alternative sources of such signals are specific processing sites, secondary structures that influence RNA degradation patterns, or chemical modifications [8–10].

The differential RNA sequencing method dRNA-seq [4] is designed to overcome these difficulties. It makes use of the 5'-monophosphate dependent terminator RNA exonuclease (TEX) that specifically degrades processed RNA, which exhibits a monophosphate at its 5' end. Transcription initiation, in contrast, produces a 5'-triphosphate that protects the unprocessed 5' end from degradation by TEX. Treating RNA isolates with TEX prior to reverse transcription to cDNA, leads to a sequencing library ([+]-library or treated library) that is enriched in primary transcription starts, compared to an untreated total RNA library ([–]-library or untreated library). Similar to other library preparation steps that enrich or deplete certain transcript types, e.g. TAP treatment [11] and rRNA depletion [12], the TEX dependent degradation of processed RNA fragments is not perfect. The [+]-library, therefore, still contains a mixture of primary and processed transcripts, albeit with a distribution of read starts that is shifted significantly towards TSS positions [4]. In the data used in this contribution a median enrichment at TSS positions of 3.5 is observable. The discrimination of TSS from other accumulations of read starts is thus non-trival and cannot be performed unambiguously from a TEX treated library alone. On the other hand comparison of [+]- and [–]-libraries offers a potentially highly informative source of information: while read starts will be relatively enriched, we can expect the alternative types of read start accumulations to be depleted in the [+]-library.

Since the signal at hand is quantitative rather than an all-or-none qualitative difference, it is imperative to employ a statistical model to assess when an observed enrichment is indeed significant. This depends strongly on the expression level. To distinguish between real TSS signals and accidental read start accumulation resulting from imperfect TEX degradation or high local expression, the aid of a background model, e.g. the [–]-library, is needed.

Hitherto, the analysis of the dRNA-seq data consists of mapping sequencing reads for each library onto the reference genome, visualizing the read coverage in a genome browser, often with displayed gene and transcription unit annotation, promoter predictions and other available prior knowledge. With this background the genome is manually inspected for positions with a more pronounced peak in the [+]- compared to the [–]-library. The interpretation of dRNA-seq signals in such a way is not only very time consuming, tedious, and error-prone, but also highly subjective and weakly reproducible. Additional annotation information from third-party sources can be very helpful but bear the risk to introduce biases, resulting in re-annotation of already "known" features, and neglecting signals that are less obviously associated with current annotation data. It is, therefore, preferable to separate dRNA-seq data analysis from subsequent data integration with additional available information.

To overcome these shortcomings we developed `TSSAR` (TSS Annotation Regime), a tool for automated *de novo* TSS annotation from dRNA-seq data. Incorporation of information like gene annotation or promoter predictions is deferred to post-processing steps.

Implementation

Theory

Detailed knowledge of the underlying background distribution is required to quantify the significance of differential read start count signals. Although related, this problem differs from the thoroughly examined problem of describing the variance in read counts per gene, which is routinely applied in the process of differential gene expression analysis. On one hand, the background is variable along the genome, depending on the transcription activity of the considered region. On the other hand, the distribution of read starts within an equally transcribed region depends on many concomitants. These are met by the different steps in the RNA-seq library construction, namely cDNA production by reverse transcriptase, fragmentation (enzymatic or mechanic), adapter ligation, read amplification by PCR, size selection, and finally the chemistry of the sequencing platform itself. Since the technology and the protocol details vary and develop with a compelling rate, it is far from trivial to capture these details [13]. Therefore, it is sensible to recollect the basic characteristic of RNA-seq data, which basically constitute count data. With this simplification we can assume that the distribution of read starts within an expressed genomic region can be modeled by a Poisson distribution with parameter λ. Given λ the Poisson probability $P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ describes the probability that k reads start at a genomic position. In dRNA-seq data genomic positions with significantly enriched differences between the Poisson distributions of [+]- and [–]-library are potential TSS. Therefore, we are concerned with finding positions where the observed difference cannot be explained easily by the local model of the background expression in the [–]-library. The difference of two Poisson distributions is given by the Skellam distribution [14] with the cumulative distribution function

$$F(D, \lambda_{[+]}, \lambda_{[-]}) = \sum_{d=-\infty}^{D} e^{-(\lambda_{[+]}+\lambda_{[-]})}\left(\frac{\lambda_{[+]}}{\lambda_{[-]}}\right)^{\frac{k}{2}} I_{|k|}\left(2\sqrt{\lambda_{[+]}\lambda_{[-]}}\right) \qquad (1)$$

Here $\lambda_{[+]}$ and $\lambda_{[-]}$ are the parameters describing the average read start rate in the [+]- and the [–]-library, respectively. $I_{|k|}$ is the modified Bessel function of the first kind and integer order $|k|$ [15].

A major practical issue is the estimation of the parameters $\lambda_{[\pm]}$ for the two libraries. We assume that read start counts per position within transcriptional active regions follow a Poisson distribution, with the expected value λ depending on the transcription rate, or to be more precise, on the RNA abundance, which depends on the transcription rate and the RNA stability. Within untranscribed regions the background, neglecting sequencing and mapping errors, ideally follows a uniform distribution with the expected value zero. Consequently, randomly selected genomic regions are most likely a mixture of transcribed and untranscribed regions. To separate the two underlying distributions and estimate the parameter λ, describing only the transcriptionally active region, a zero-inflated Poisson model regression [16, 17] is applied. For each sample Y the probability ϕ that an observed zero is a structural zero (i.e., part of a transcriptional inactive region and thus from a uniform zero distribution) and not part of the transcriptional active region is estimated, such that

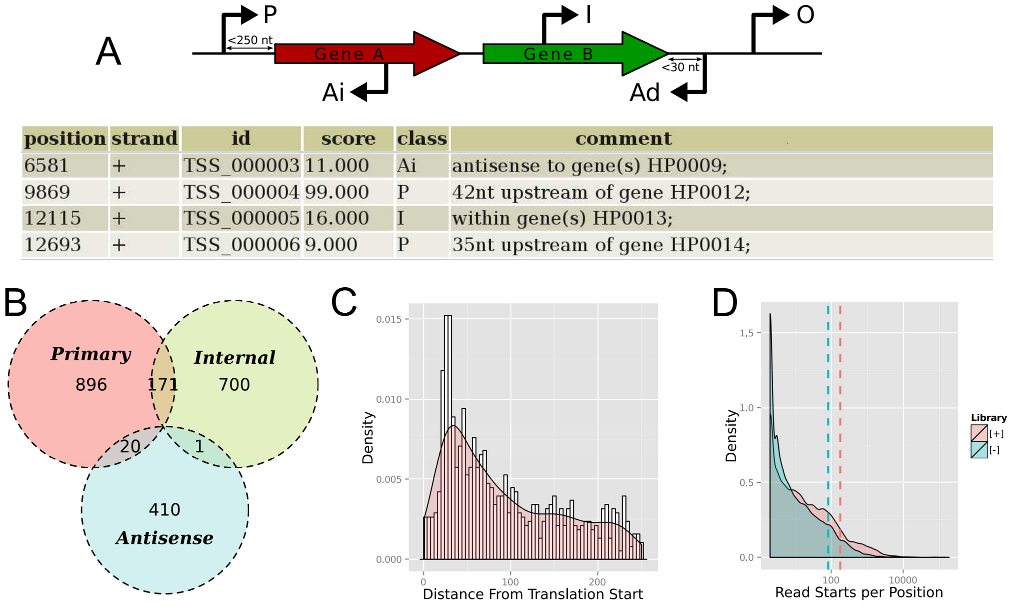
$$P(Y = 0) = \phi + (1 - \phi) \cdot e^{-\lambda} \tag{2}$$

where $e^{-\lambda}$ is the probability for a position within the Poisson distributed part to have zero reads starting there (sampling zero). These positions are part of transcriptional active regions. We use a zero-inflated Poisson regression to estimate ϕ and thus determine how many positions without read starts are structural and sampling zeros, respectively. Only the latter and positions that have at least one read start are used to estimate λ of the [+]- and [–]-library, respectively. The estimation of λ thus effectively considers the transcriptionally active regions only.

Program architecture

`TSSAR` has been implemented in `Perl` and `R` and is available in two variants: A stand-alone version incorporates the core statistic routines and is best suited to be used in custom high-throughput dRNA-seq analysis. The Web service (available at http://rna.tbi.univie.ac.at/TSSAR/) comprises additional components for pre- and post-processing, thus providing a Web-based, cross-platform compatible pipeline for dRNA-seq analysis. An overview of the pipeline workflow can be found in supplemental Figure 1.

The **TSSAR Web service** is built on top of the `Perl Dancer` [18] framework and adheres to the Representational State Transfer (REST) [19] principles of Web architecture. The first step in using the `TSSAR` online pipeline is pre-processing of mapped reads, i.e., extracting the essential information of read start counts per genomic position. To avoid the necessity of uploading huge mapping files (typically for bacterial genomes up to several gigabytes), we implemented the **TSSAR client** for local pre-processing of mapped reads in SAM/BAM or BED format on the user's computer. To grant platform independence, the `TSSAR` client is implemented in `Java`. Once the relevant data is extracted from the mapping files assisted by the `Picard tools` [20], files are compressed using `XZ utils` [21] and automatically transferred, using the `Apache HttpComponents` [22] package, to the `TSSAR` Web server. On the Web server the statistical calculations are conducted and potential TSS are predicted. The `TSSAR` Web service provides an assortment of post-processing steps. The list of predicted TSS can be reduced by merging consecutive TSS and cluster them into the most prominent position. For samples where the reference genome annotation was specified, all annotated TSS are classified into primary, internal, anti-sense or orphan, according to their position relative to nearby genes, see Figure 1A. Based on the classification the 5' UTR length distribution is determined. All results are visualized and provided for download. Figure 1 depicts partly the output for showcase data sets [4, 23]. Beside the shown results, the output additionally contains all annotated TSS and the clustered



position	strand	id	score	class	comment
6581	+	TSS_000003	11.000	Ai	antisense to gene(s) HP0009;
9869	+	TSS_000004	99.000	P	42nt upstream of gene HP0012;
12115	+	TSS_000005	16.000	I	within gene(s) HP0013;
12693	+	TSS_000006	9.000	P	35nt upstream of gene HP0014;

Figure 1: **Post-processing and Visualization.** (A) Similar, but more restrictive, to the scheme in [4] each annotated transcription start site is classified according to its genomic context: If a TSS is positioned within 250 nt upstream of an annotated gene, it is classified as **P**rimary. TSS within an annotated gene is labeled **I**nternal. A TSS which is on the opposite strand of an annotated gene is classified as **A**ntisense. This class further splits into **Ai** and **Ad**, for internal antisense and downstream antisense, respectively. The latter is reserved for a TSS which points in the opposite reading direction and is less than 30 nt downstream of an annotated gene. A TSS that falls in none of these classes is reported to be **O**rphan. (B) As a matter of fact, one TSS can have several labels as it might fall into more than one of the aforementioned classes. The `TSSAR` Web service summarizes the counts of the overlapping main classes graphically. (C) For TSS which are annotated as 'Primary' the 5'UTR lengths are deduced and the corresponding distribution is plotted. (D) To assess the efficiency of the TEX treatment, the distribution of read starts per position is provided as a helpful indicator. If the enrichment in the [+]-library worked efficiently, we expect fewer read start sites, each of which will have more reads. Hence the distribution is flattened on the left side and bulged at the right side. The corresponding distribution and the mean (dashed line) is expected to be shifted to the right compared to the [–]-library.

TSS list in BED [24] and GFF format. All tables are available in comma and tab-separated lists, as excel and HTML files. With the assistance of the pre-computed plots, it is easy to gain a quick overview of the quality of the analysis.

While the `TSSAR` Web service provides convenient usability for routine dRNA-seq analysis tasks, there is also a demand for integrating third-party bioinformatics tools into custom analysis pipelines. To address this issue, we provide a **TSSAR stand-alone version**. In this version, the implementation is restricted to processing of SAM files, analysis based on the statistical calculations, and output of annotated TSS in BED format. The stand-alone version is available for download from the `TSSAR` Web site.

Statistical calculation

We chose a sliding window approach with a dynamic assessment of each position in the context of its local surrounding in order to account for different transcription rates across the genome. As a matter of fact, the choice of the window size parameter has an effect on the results (see supplementary Figure 2). There, two conflicting interests have to be balanced. On the one hand, the region should be large enough to provide enough information for a reliable distribution parameter estimation. On the other hand, the region should be small enough to provide an as homogeneous surrounding as possible. If the sliding window covers more than one actively transcribed gene, with different RNA abundances, the variance will be estimated over all transcribed entities. This might blur small signals, e.g., for low abundant sRNA genes. As a compromise, the default window size is 1,000 nt, approximately matching the average length of prokaryotic genes. It can be easily adjusted by the user.

For each window the parameters describing the Poisson distribution are estimated in the following manner: First, the sample values are winsorized [25], i.e., the highest read start count is substituted with the second highest count. The same procedure is done for the lowest value. This increases the robustness of the method against outliers, which may be caused by mis-mapping and/or abundant RNA fragments e.g. arising from rRNA loci.

Second, the zero-inflated Poisson regression is applied to estimate ϕ, the probability that an observed zero is a structural zero from an untranscribed region instead of a sampling zero from a transcribed region. The `R` package `VGAM` is used for the regression [17, 26]. Here, the parameters describing the Poisson distribution are fitted by full maximum likelihood estimation (MLE). In case the MLE algorithm fails to converge, which might happen because the underlying assumption of a well behaved Poisson distribution is violated, the respective window is excluded from further analysis. While this might seem to be a drawback, it serves

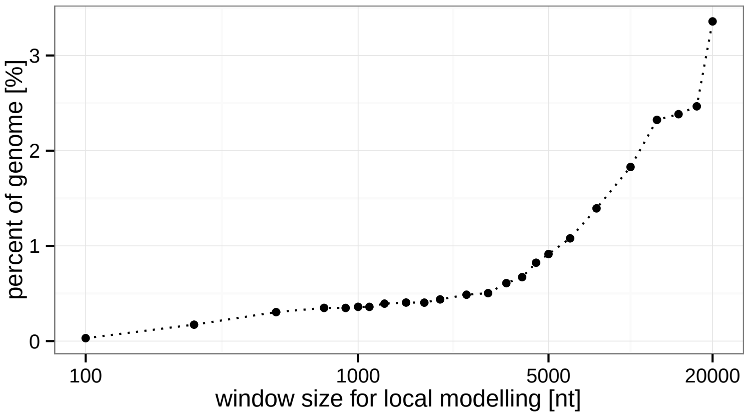


Figure 2: **Regions of non-convergence.** Regions where the applied zero-inflated Poisson regression does not converge are omitted from the analysis and need manual inspection. Since the basic unit which cannot converge is the step size (equals a tenth part of the windows size) there is a correlation between the parameter window size and the percentage of the genome which can not be modeled. The *H. pylori* dRNA-seq data (see section Evaluation) shows that for all practical useful window sizes below 5,000 nt, less then 1% of the genome eludes analysis.

as a minimal plausibility check, ensuring the data fulfills the underlying assumption of following a Poisson distribution. Sequencing libraries with low complexity but many PCR duplicates might otherwise feign confidence in the results, which can actually not be deduced from the data. A BED file listing the omitted segments which typically correspond to less than 1% of the genome is provided (see Figure 2). In the evaluation data set (see section Evaluation) modeled with a window size of 1,000, 24 regions with a total length of 12,000 bases could not be modeled (∼0.5% of the genome). The majority correspond to tRNA and rRNA coding loci (10 and 5 single regions, respectively). Additionally, 4 regions overlapped with annotated protein coding genes and the remaining 5 did not overlap with any annotated gene. A manual screening of the corresponding regions revealed that they share common characteristics. Generally they are small islands with very high expression levels.

Third, a regression procedure is applied to each window in the [+]- and in the [–]-library separately. For each library the probability ϕ is transformed into an expected number of excess structural zeros. Since the same genomic region is under consideration in both libraries, a similar proportion of untranscribed and transcribed regions can be expected. To increase robustness, the average between the number of structural zeros in both libraries is calculated and the estimated number of zeros are removed from each library. To determine λ for each library, describing the Poisson distribution of the sample, the arithmetic mean of the remaining counts is calculated.

In the next step the probability that the read start differences between [+]- and [–]-library can be explained by the aforementioned background model is calculated. For this purpose, the original read start counts are normalized by

$$\widehat{p}_i = \begin{cases} p_i \cdot \frac{\sum M}{\sum P} & , \sum M > \sum P \\ p_i \cdot 1 & , \sum M \leq \sum P \end{cases} \tag{3}$$

$$\widehat{m}_i = \begin{cases} m_i \cdot \frac{\sum P}{\sum M} & , \sum P > \sum M \\ m_i \cdot 1 & , \sum P \leq \sum M \end{cases} \tag{4}$$

Thereby, p_i and $\widehat{p}_i$ are the raw and normalized values of the [+]-library at position i, respectively. $\sum P$ and $\sum M$ are the native sums of all read start counts in the total [+]- and [–]-library, respectively. The same applies to the [–]-library, i.e., m_i and $\widehat{m}_i$. The effect of this step is to scale the read counts of the larger library relative to the smaller one, hence avoiding artificial distending of the sample variance. The estimated parameters $\lambda_{[+]}$ and $\lambda_{[-]}$ are therefore normalized accordingly.

For each sequence position i in the current window, the difference $\widehat{d}_i = \widehat{p}_i - \widehat{m}_i$ of the normalized counts between [+]- and [–]-library is calculated. Unexpectedly large positive values of $\widehat{d}_i$ for position i indicate TSS, while exceptional negative values may indicate processing sites. The probability of observing $\widehat{d}_i$ is evaluated w.r.t. the Skellam distribution with the estimated normalized Poisson parameters.

The window slides along the genome with a step size equal to $1/10^{th}$ of the window size, hence each position is evaluated in 10 slightly different contexts. The geometric mean of all ten p-values is calculated in order to obtain the final position-wise p-value. Finally, each position that falls below a user-specified average p-value cutoff and whose total read start count in the [+]-library exceeds a user specified noise cutoff is reported as a significant TSS. The noise cutoff serves as an additional safeguard to restrict the results to plausible annotations. This is needed because the Skellam distribution works only with the differences of the expectation values of the underlying Poisson distributions. A very high expectation value in the [–]-library in combination with a small expectation value in the [+]-library leads to a negative expectation value of the resulting Skellam distribution. This, in turn could lead to annotated positions which are not supported by reads in the [+]-library, as significantly enriched. To prevent this unwanted behaviour a user defined number of read starts must be observed in the [+]-library.

Results

The goal of the `TSSAR` method is to provide user-friendly tools for rapid annotation of significant TSS based on dRNA-seq data. We therefore implemented a stand-alone version and a Web service. The first is intended to be used in high-throughput analysis pipelines whereas the latter represents an easy to use and platform independent user interface. For a Web service it is important to avoid the transfer and storage of gigabyte-sized mapping files. We therefore provide a `Java` client that extracts the necessary information and asks the user for only two parameters, namely genome size and window size. The data is pre-processed locally on the user's computer. The essential information, i.e., the number of sequencing reads starting at each position, is automatically uploaded and analyzed on the `TSSAR` Web server. All relevant cutoffs like p-value and noise threshold are subsequently selectable for precomputed values.

Evaluation

To evaluate the performance of `TSSAR` in analyzing dRNA-seq data, we resort to the published data set for *Helicobacter pylori* [4]. We used the publicly available raw sequencing data from the Sequence Read Archive [27] (study accession number SRP001481), restricting ourselves to the dRNA-seq data from mid-logarithmic growth phase and acid stress growth condition. The reads were pooled and mapped to the reference genome (NCBI accession ID NC_000915) using `segemehl` version 0.1.4 [28] with default parameters.

Based on this data, which were normalized in the same way as indicated in equations 3 and 4, we predicted putative TSS with three different approaches. The first two represent a naïve benchmark. First, we calculated the difference $(\widehat{p}_i - \widehat{m}_i)$ for each position i of the [+]- and [–]-library read start counts. We applied a different cutoff threshold between 1 and 300, thereby denoting every position with a difference higher than the cutoff to be a putative TSS. The resulting list of potential TSS was compared to the manual annotation from [4] using `BEDTools Intersect` [29], allowing ±2 nt inaccuracy to call a manual and an automated annotated TSS the same. The second approach is quotient based. Analogous to the difference based approach, the quotient $\frac{\widehat{p}_i+1}{\widehat{m}_i+1}$ is calculated for each position i (+1 is used as pseudo-count to avoid division by zero). Again we use different cutoff values between 1.1 and 20. These two approaches have their static nature in common. The same threshold is applied for the whole genome. A similar approach was already applied by [30]. Albeit, there it was used to identify differentially induced TSS between different strains and growth conditions and additional information about promoter sequences was used to gain specificity.

Finally, we applied the dynamic `TSSAR` model, which analyzes the transcriptome locally and thus is able to model the different dynamics within the transcriptome. Here, we used a window size of 1,000 nt (approximately the mean gene length in *H. pylori*) and a noise cutoff of 3 reads per position. We filtered with different p-value threshold from $1 \cdot 10^{-15}$ to $9 \cdot 10^{-1}$.

From these results, each threshold based prediction is evaluated using standard measurements: recall rate ($\frac{TP}{TP+FN}$), precision ($\frac{TP}{TP+FP}$), accuracy ($\frac{TP+TN}{TP+FP+FN+TN}$) and the F-measure ($2 \times \frac{precision \times recall}{precision+recall}$) [31], where TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively. Figure 3 depicts the results of this analysis. `TSSAR` shows a much higher precision and simultaneously a less sharp decrease of the recall rate. In terms of the F-measure, it outperforms the fixed-threshold approaches by about 2-fold. A further major advantage is the smoother course of the F-measure along different p-value cutoffs. This makes the resulting annotation less dependent on the cutoff choice. The optimal cutoff value for the basic annotation strategies based on difference or ratio might be very variable for different experiments and difficult to deduce without a reference annotation.

In its default settings `TSSAR` merges consecutive TSS. Since the tested naïve approaches do not share this behavior, we tested the influence of TSS clustering on the prediction performance separately (see supplementary Figure 5). Although, clustering contributes to the precision of the prediction, the effect is much too small to cause the improved performance of `TSSAR`.

Additionally, besides comparing our automated annotation to the manual annotation by the authors, we examined how precise `TSSAR` reproduces known *H. pylori* TSS. Therefore, we used TSS studied in detail by independent methods, such as primer extension or 5' RACE. From the 74 examples described in the literature (summarized in supplementary material of [4]), we calculated the distance to the closest position which we annotated as TSS. If the discrepancy was more then 10 nt, we considered the TSS as not recovered. Figure 4 shows the result of this analysis for two `TSSAR` annotations with different parameters. The first one with lenient threshold values (aiming for sensitivity), and the later with more stringent values (aiming for specificity). In both cases the majority of experimentally confirmed TSS could be detected at the exact same position (39 and 37 TSS, respectively). `TSSAR` missed 14 and 21 TSS, respectively, compared to the 12 TSS that were also not detectable in the manual annotation by the authors of [4]. We have to emphasis that, in contrast to a manual annotation, our method is not aware of any annotation information, which might induce a human curator to prefer certain positions. Comparison of the two naïve approaches and TSSAR emphasises that the presented statistical method is relatively insensetive to certain parameter thresholds, see supplementary Figure 3.

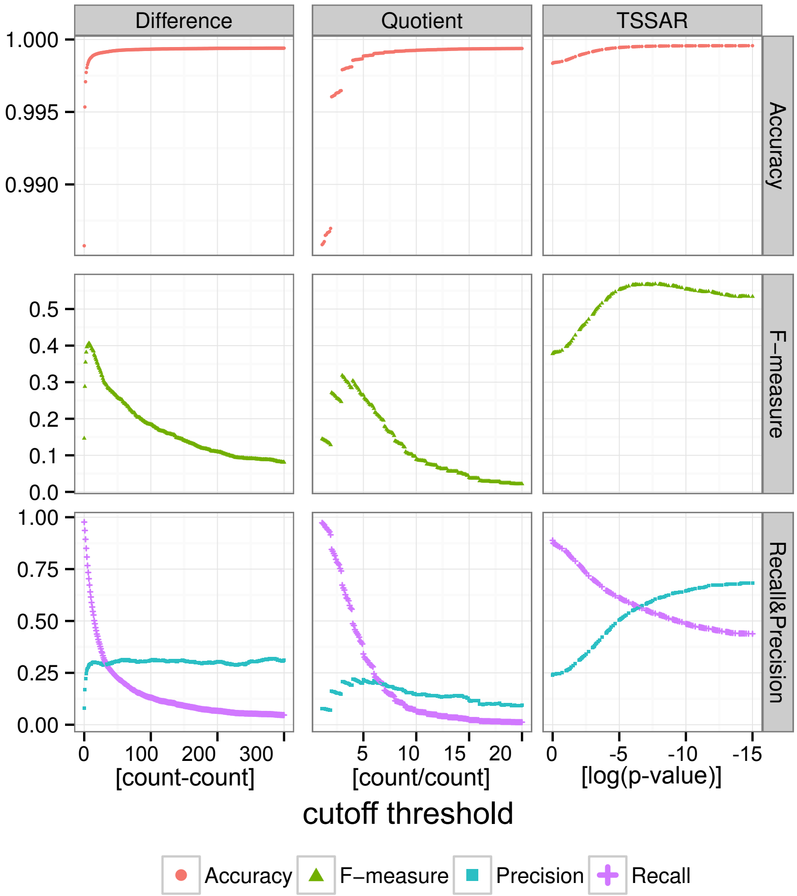


Figure 3: **Evaluation of `TSSAR` performance.** Comparison of the prediction power of `TSSAR` against two fixed-cutoff approaches *Difference* and *Quotient*. For each method different cutoff thresholds were applied. The difference, quotient and logarithm of the p-value are plotted along the x-axis. Please note, for comparability the log(p-value) is plotted in descending order from left to right. The resulting predictions were evaluated by calculating the recall rate, precision, F-measure and accuracy. The dynamic approach of `TSSAR` clearly outperforms the remaining in all aspects. Since only `TSSAR` applies a clustering of consecutive TSS positions, this effect was separately examined, results can be found in supplementary Figure 5.

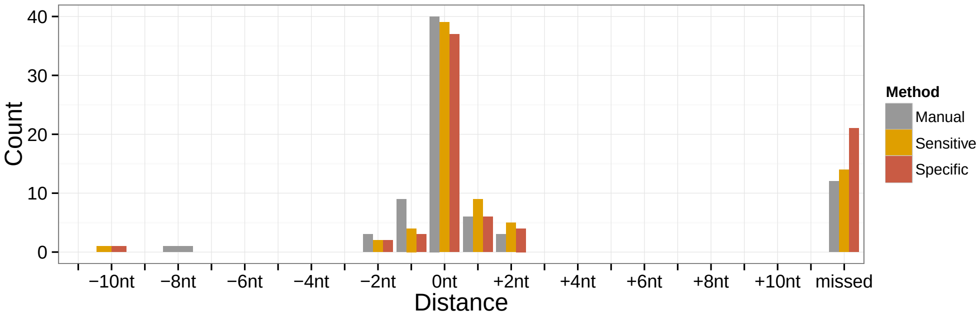


Figure 4: **Recall experimental validated TSS.** Comparison of 74 experimentally validated TSS described in literature [4] with `TSSAR` results. The *Manual* TSS annotation recovered 40, 15 and 6 TSS with a 0, ±1 and ±2 nt offset, respectively. Here 12 TSS were annotated more than 10 nt away from the experimentally determined position (summarized as *missed* in the plot). `TSSAR` was run with a *Sensitive* and a *Specific* parameter set (p-value cutoff 0.05 and 0.0001; noise cutoff 1 and 3, respectively). With sensitive parameters 39 TSS (53%) were annotated on the exact same position. Of the remaining TSS 13 and 7 were annotated with ±1 and ±2 nt variance, respectively, whereas 14 TSS (19%) were annotated more than 10 nt away. The specific `TSSAR` prediction annotated 37, 9 and 6 TSS with 0, ±1 and ±2 nt offset, respectively, relative to the experimentally validated position. In this case 21 TSS (28%) were annotated more than 10 nt away, and therefore annotated as missed. The results of the same analysis including also our naïve benchmark approaches can be found in supplementary Figure 3.

Discussion

A major advantage of an automated TSS annotation, based on a sound statistical analysis, neglecting *a priori* knowledge of the whereabouts of promoters and other already established annotation, lies in the avoidance of any bias towards certain genomic positions. This ensures an unbiased analysis as well as a comparable and reproducible TSS annotation procedure.

Although our approach checks whether the basic assumption that the read starts of a sequencing library are Poisson distributed holds, a manual inspection of the produced data is still recommended. The automated TSS prediction is only as good as the underlying dRNA-seq libraries. We therefore emphasize that a thoughtful investigation of the input sequencing reads, especially for PCR duplicates, is advised. Manual inspection is necessary for those genomic regions that are not annotated by `TSSAR` due to non-convergence in the estimation of the expression parameters. For `TSSAR`'s output, we recommend at least a basic sanity check, since very complex regions, such as tRNA and rRNA loci, might be misconstrued. In spite of these precautions, the work load to check hundreds or a few thousands of predicted TSS positions is significantly reduced compared to screening millions of genomic positions in the first place.

Reliable and automated TSS annotation is a prerequisite for many applications. So far, most genome-wide TSS annotations focused on a static picture of the transcriptomal architecture [2, 32] (there are also notable exceptions, e.g. [30, 33]). One reason is that data analysis was more laborious than data generation. Relieving the experimenter from this time-consuming burden might liberate the resources to investigate more of the dynamics and alteration of the transcriptome, due to external stimuli or evolutionary differences. During manuscript preparation the latter was demonstrated by conducting a comparative transcriptomics approach [34]. There, TSS annotation was also conducted in an automated manner. First, putative TSS are selected by considering the "flank height", and the differences of mapped read starts of position $i-1$ to i are calculated. These sites are then evaluated similarly to our *Quotient* approach based on the ratio between the TEX treated and untreated library. The problem of selecting an educated cutoff, which is immanent to all methods but especially troublesome for classifiers which directly depend on variable conditions such as sequencing depth, was neatly circumvented by using a comparative approach. Transcriptomes of different *Campylobacter jejuni* isolates were used to dynamically adjust thresholds if signals in different strains could be observed. In the more typical application scenarios, where such comparative information is not available, a robust p-value estimate that takes the dynamic range of transcription activity along the whole genome into account for the classification seems to be preferable.

Currently, `TSSAR` is based on the assumption that a [+]- and [–]-library is analyzed and only positions with a significant enrichment in the [+]-library are reported as potential TSS. At least two other application scenarios of the statistical framework are possible. One is to detect RNA processing sites and the other to analyze differentially induced transcription starts. In principle the latter could be achieved by comparing two TEX treated libraries resulting from dRNA-seq runs of different growth conditions. In that case, a large positive and negative $\widehat{d_i}$ is of importance as it indicates (growth phase dependent) induction of a TSS in the one or the other library. RNA processing sites are in principle detectable using the "standard" dRNA-seq approach. Positions where a significant enrichment in the [–]-minus over the [+]-library is observable are of interest. Extremely small values of $\widehat{d_i}$ point to these positions. Tackling both issues, processing sites and induced transcription initiation, is however currently hampered by the lack of experimentally verified training sets. Furthermore, although tailored for analyzing dRNA-seq data, in principle, the `TSSAR` method should be applicable to other RNA-seq protocols, e.g., [11], which aim to enrich read starts at certain positions in the sequencing library. Currently, the run-time of `TSSAR`, see supplementary Figure 4, prevents its application for one of the above mentioned purposes to complete eukaryotic genomes. An improvement of this aspect will be a task for the future development and refinement of the program.

The modularity of the `TSSAR` framework makes it possible to extend the current approach e.g., by improving the statistical model. Alternative approaches based on a different (non-Poisson) distribution or the Pitman sampling method [6] can be implemented in the `TSSAR` core module, without the necessity to change the `Java` client or the Web service front end. The RESTful architecture of the `TSSAR` Web service provides additional extensibility, rendering implementation of new functionality such as promoter or operon characterization straightforward.

Conclusion

Here, we presented an automated analysis of dRNA-seq data which aims to detect significantly enriched TSS positions. The background distributions of sequencing read starts are modeled locally by a zero inflated Poisson distribution. Positions with a larger difference between the TEX treated and the untreated library than expected, considering the background, are annotated as significant transcription start sites. We could show that our method reproduces manually analyzed dRNA-seq data better than two simple approaches that use a global cutoff to discriminate between true and false signals. Furthermore, the choice of a *p*-value cutoff is more intuitive and less arbitrary.

`TSSAR` is available both as a stand alone tool and as a Web service at http://rna.tbi.univie.ac.at/TSSAR/. The latter provides additional post-processing functionality like TSS classification or merging of consecutive TSS. The `TSSAR` Web service offers user-friendly and intuitive online access to the `TSSAR` framework whereas the stand-alone version is intended for integration into third-party annotation pipelines.

Availability and requirements

- **Project name:** `TSSAR`
- **Project home page:** http://rna.tbi.univie.ac.at/TSSAR
- **Operating system:** Platform independent
- **Programming language:** `Java`, `Perl` and `R`
- **Other requirements:** Client needs `Java` 1.6 or higher and the standalone version is based on `Perl` 5, R 2.15
- **License:** `Java` client under Apache License, Statistics module under GPL2.
- **Any restrictions to use by non-academics:** For non-profit use only.

Competing interests

None.

Author's contributions

FA implemented the statistical analysis and evaluated the performance, SF programmed the `Java` client, MTW, FA and RL implemented the Web service. All authors contributed to the implementation details and testing, collaborated in writing and approved the final manuscript.

Acknowledgments

We thank David Hot for initiating our interest in dRNA-seq and sharing data, Thomas Yee for valuable advice and Peter Kerpedjiev for proof reading.

Funding This work was partly funded by the Austrian Science Fund (FWF) projects "In silico annotation of noncoding RNAs and their targets" (AF 0430511), "RNA regulation of the transcriptome" (F43), the German ministry of science (0316165C as part of the e:Bio initiative), and the Deutsche Forschungsgemeinschaft (STA 850/10-2 within the SPP 1258 "Sensory and regulatory RNAs in Prokaryotes").

References

1. Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using RNA-seq**. *Current opinion in microbiology* 2010, **13**(5):619–624.
2. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ: **The transcription unit architecture of the Escherichia coli genome**. *Nature biotechnology* 2009, **27**(11):1043–1049.
3. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons B, Sorek R: **A single-base resolution map of an archaeal transcriptome**. *Genome research* 2010, **20**:133–141.
4. Sharma C, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, et al.: **The primary transcriptome of the major human pathogen Helicobacter pylori**. *Nature* 2010, **464**(7286):250–255.
5. Schmidtke C, Findeiß S, Sharma C, Kuhfuß J, Hoffmann S, Vogel J, Stadler P, Bonas U: **Genome-wide transcriptome analysis of the plant pathogen Xanthomonas identifies sRNAs with putative virulence functions**. *Nucleic acids research* 2012, **40**(5):2020–2031.
6. Tauber S, von Haeseler A: **Exploring the sampling universe of RNA-seq**. *Statistical applications in genetics and molecular biology* 2013, **12**(2):175–188.
7. Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH: **Structure and complexity of a bacterial transcriptome**. *Journal of bacteriology* 2009, **191**(10):3203–3211.
8. Ishitani R, Yokoyama S, Nureki O: **Structure, dynamics, and function of RNA modification enzymes.** *Current Opinion in Structural Biology* 2008, **18**(3).
9. Knoop V: **When you can't trust the DNA: RNA editing changes transcript sequences.** *Cellular and Molecular Life Sciences* 2011, **68**(4).
10. Findeiß S, Langenberger D, Stadler PF, Hoffmann S: **Traces of post-transcriptional RNA modifications in deep sequencing data**. *Biological Chemistry* 2011, **392**(4).
11. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, Archambaud C, Cossart P, Sorek R: **Comparative transcriptomics of pathogenic and non-pathogenic Listeria species**. *Molecular systems biology* 2012, **8**.
12. Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, Livny J, Earl AM, Gevers D, Ward DV, Nusbaum C, Birren BW, Gnirke A: **Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes.** *Genome Bilogogy* 2012, **13**(r23).
13. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M: **Modelling and simulating generic RNA-Seq experiments with the flux simulator**. *Nucleic acids research* 2012, **40**(20):10073–10083.
14. Skellam J: **The frequency distribution of the difference between two Poisson variates belonging to different populations.** *Journal of the Royal Statistical Society. Series A (General)* 1946, **109**(Pt 3):296.
15. Abramowitz M, Stegun IA: **Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing.** 1972. [Modified Bessel functions of the first kind with integer coeficient appear frequently in many fields of physics and as the result of integrals in mathematical statistics. They rapidly grow as a function of their argument.].
16. Lambert D: **Zero-inflated Poisson regression, with an application to defects in manufacturing**. *Technometrics* 1992, **34**:1–14.
17. Yee TW: **The VGAM package for categorical data analysis**. *Journal of Statistical Software* 2010, **32**(10):1–34.
18. **The Perl Dancer Project**[http://www.perldancer.org].
19. Fielding RT: **REST: Architectural Styles and the Design of Network-based Software Architectures**. *Doctoral dissertation*, University of California, Irvine 2000.
20. **Picard tools version 1.85**[http://picard.sourceforge.net].
21. **XZ for Java version 1.2**[http://tukaani.org/xz/java.html].
22. **Apache HttpComponents version 4.2.3**[http://hc.apache.org/index.html].

23. Ramachandran V, Shearer N, Jacob J, Sharma C, Thompson A: **The architecture and ppGpp-dependent expression of the primary transcriptome of Salmonella Typhimurium during invasion gene expression**. *BMC genomics* 2012, **13**:25.

24. Quinlan AR, Hall IM: **BEDTools-User-Manual**[bedtools.googlecode.com/files/BEDTools-User-Manual.pdf].

25. Searls DT: **An estimator for a population mean which reduces the effect of large true observations**. *Journal of the American Statistical Association* 1966, **61**(316):1200–1204.

26. Yee TW, Yee MT, Suggests M: **Package 'VGAM'** 2012.

27. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al.: **Database resources of the national center for biotechnology information**. *Nucleic acids research* 2012, **40**(D1):D13–D25.

28. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J: **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. *PLoS computational biology* 2009, **5**(9):e1000502.

29. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841–842.

30. Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM: **Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120**. *Proceedings of the National Academy of Sciences* 2011, **108**(50):20130–20135.

31. Sokolova M, Japkowicz N, Szpakowicz S: **Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation**. In *AI 2006: Advances in Artificial Intelligence*, Springer 2006:1015–1021.

32. Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voß B, Steglich C, Wilde A, Vogel J, et al.: **An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803**. *Proceedings of the National Academy of Sciences* 2011, **108**(5):2124–2129.

33. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, et al.: **Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis**. *Science* 2012, **335**(6072):1103–1106.

34. Dugar G, Herbig A, Förstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM: **High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple Campylobacter jejuni Isolates**. *PLoS genetics* 2013, **9**(5):e1003495.

8 Target prediction using conservation information

8.1 Statement of personal contribution

HT, PFS and ILH designed the study. HT implemented the program and wrote the paper. HT, FE, FA contributed to the evaluation of the program for the alignment and single sequence mode.

8.2 Article

Hakim Tafer, Fabian Amman, Florian Eggenhofer, Peter F. Stadler, and Ivo L. Hofacker.
"Fast accessibility-based prediction of RNA–RNA interactions."
Bioinformatics 27, no. 14 (2011): 1934-1940.

Received on September 9, 2010,
Revised on April 28, 2011,
Accepted on April 29, 2011;

© 2010 by the authors; licensee Oxford University Press; Reprint with permission from Oxford University Press, and the authors.

BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 14 2011, pages 1934–1940
doi:10.1093/bioinformatics/btr281

Structural bioinformatics
Advance Access publication May 18, 2011

Fast accessibility-based prediction of RNA–RNA interactions

Hakim Tafer[1,*], Fabian Amman[2], Florian Eggenhofer[2], Peter F. Stadler[1,2,3,4,5] and Ivo L. Hofacker[2,*]

[1]Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany, [2]Institute for Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria, [3]Max Planck Institute for Mathematics in the Sciences, [4]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, D-04103 Leipzig, Germany and [5]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM-87501, USA

Associate Editor: Martin Bishop

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

ABSTRACT

Motivation: Currently, the best RNA–RNA interaction prediction tools are based on approaches that consider both the inter- and intramolecular interactions of hybridizing RNAs. While accurate, these methods are too slow and memory-hungry to be employed in genome-wide RNA target scans. Alternative methods neglecting intramolecular structures are fast enough for genome-wide applications, but are too inaccurate to be of much practical use.

Results: A new approach for RNA–RNA interaction was developed, with a prediction accuracy that is similar to that of algorithms that explicitly consider intramolecular structures, but running at least three orders of magnitude faster than `RNAup`. This is achieved by using a combination of precomputed accessibility profiles with an approximate energy model. This approach is implemented in the new version of `RNAplex`. The software also provides a variant using multiple sequences alignments as input, resulting in a further increase in specificity.

Availability: `RNAplex` is available at www.bioinf.uni-leipzig.de/Software/RNAplex.

Contact: htafer@bioinf.uni-leipzig.de; ivo@tbi.univie.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* Online.

Received on September 9, 2010; revised on April 28, 2011; accepted on April 29, 2011

1 INTRODUCTION

The status of RNA in molecular biology has changed dramatically over the last decade. Instead of taking on a rather marginal role as messenger of genomic information, they are now considered as key regulatory elements in a wide spectrum of cellular processes. As of 2008, the number of known non-coding RNA sequences reached an overwhelming 29 million grouped into 1300 distinct families (Gardner *et al.*, 2009).

Non-coding RNAs (ncRNAs) frequently function by binding to other RNAs. For example, snoRNAs mediate pseudouridylation and methylation of rRNAs and snRNAs (Bachellerie *et al.*, 2002) and can influence the splicing of pre-mRNAs (Zorio *et al.*, 1997). ncRNAs are also involved in the editing of other RNA sequences (Benne, 1992), transcription and translation control (siRNA, miRNA, stRNA) (Banerjee and Slack, 2002; Fire *et al.*, 1998; Kugel and Goodrich, 2007) or plasmid replication control (Eguchi and Tomizawa, 1990). While siRNAs are often fully complementary to their targets, most other ncRNAs interact in a more intricate manner, which does not involve perfect hybridization. For example in *Escherichia coli.*, *OxyS*, which is involved in oxidative stress response, interacts with its target mRNA, *fhlA*, through formation of a two sites kissing complex (Argaman and Altuvia, 2000). Although there is statistical evidence that a plethora of ncRNAs interacts with other RNAs (The Athanasius F. Bompfünewerer RNA Consortium: *et al.*, 2007), targets remain unknown for most of them. The prediction of RNA–RNA interactions, therefore, has become an important field in computational biology.

RNA–RNA interactions are primarily governed by the same types of hydrogen bonds and stacking interactions as RNA secondary structure formation. The problem can, therefore, be tackled by similar algorithmic approaches and the same parametrization of the interaction energies. We may distinguish two distinct ways of addressing the RNA–RNA interaction problem. The most straightforward way consists in concatenating both sequences and subsequently folding them as a pseudo-single sequence. The precision of this kind of approach depends greatly on how the concatenation is handled. The crudest approaches use linker sequences to connect both RNA strands (Stark *et al.*, 2003). This can lead to erroneous structure prediction as the linker may interfere with the interacting sequences. Alternatively, a small modification of the folding algorithm keeps track of the concatenation point(s) and uses adjusted energy parameters for the loops in which the junctions occur (Andronescu *et al.*, 2003; Bernhart *et al.*, 2006b; Dimitrov and Zuker, 2004; Dirks *et al.*, 2007; Hofacker *et al.*, 1994). A combinatorially different model, known as RNA–RNA interaction problem (RIP), covers a larger set of possible structures (Alkan *et al.*, 2006; Chitsaz *et al.*, 2009; Huang *et al.*, 2010; Pervouchine, 2004).

The second type of approaches conceptually decomposes the RNA hybridization process into two stages: (i) the unfolding of the interacting regions of the two partners and (ii) the direct interaction of the exposed binding sites. In practice, one first computes the probability of being unpaired for each region (sequence interval) in both sequences. These probabilities are equivalent to the free energy necessary to expose the regions. In the second step, the interaction energy between each combinations of regions is evaluated (Busch *et al.*, 2008; Mückstein *et al.*, 2006, 2008). This approach was,

---

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

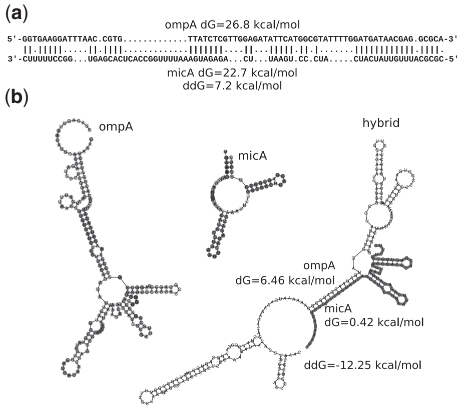


Fig. 1. Comparison of the ompA–micA hybrids predicted with and without considering intramolecular structures. (**a**) Hybrid structure predicted with `RNAplex` without considering the intramolecular structures of the RNA sequences. The hybrid extends over 67 and 69 nucleotides on ompA and micA, respectively, and has an hybridization energy of -42.3 kcal/mol. Still the energy needed to unfold both binding regions on ompA and micA amounts $22.7+26.8=49.5$ kcal/mol, larger than the energy gained through binding. (**b**) ompA–micA interaction predicted by `RNAup`. OmpA–micA hybrid is shown on the right hand side, with the micA sequence represented by a bold line. Even though the hybrid is much smaller than the interaction in (**a**), it has a lower total interaction energy (ddG) of -12.25 kcal/mol, due to the fact that the interacting regions are less structured.

in particular, applied successfully to sRNA–mRNA interactions in bacteria.

While both types of algorithms proved useful in predicting the correct interaction structure of a ncRNA with its (known) target, they are computationally expensive, requiring at least $\mathcal{O}((n+m)^3)$ operations, where n and m are the size of the target and query sequences, respectively, and hence are impractical for genome-wide target predictions.

A drastic reduction in computational complexity can be achieved by omitting the computation of secondary structures within the monomers, as demonstrated by `RNAhybrid` (Rehmsmeier *et al.*, 2004), which runs in $\mathcal{O}(m\cdot n\cdot L^2)$ when restricting the maximum loop length to L. `RNAplex`, a conceptually very similar approach (Tafer and Hofacker, 2008), further reduces the time complexity to $\mathcal{O}(m\cdot n)$ by using a modified energy model. Neglecting the internal structure of the interacting sequences leads to a drastic decrease in specificity; however, see Figure 1. This issue is roughly addressed by `RNAplex` in that it mimics the effect of the competition between intra- and intermolecular interactions by adding a fixed per-nucleotide penalty (Tafer and Hofacker, 2008).

Currently, one therefore has to choose between precise but impractically slow methods or fast but imprecise methods for ncRNA target search, a situation that is quite unsatisfactory. In this contribution, we extend the `RNAplex` approach Tafer and Hofacker (2008) to tackle this problem. We mimic the effect of the competition between intra- and intermolecular interactions by adding a position-dependent per-nucleotide penalty instead of a fixed penalty. This penalty is derived from precomputed accessibility profiles produced by `RNAplfold` (Bernhart *et al.*, 2006a; Bompfünewerer *et al.*, 2008a) or `RNAup` (Mückstein *et al.*, 2008). More explicitly, these profiles contain the probabilities that any subsequence of arbitrary length is unpaired in thermodynamic equilibrium. These probabilities are converted to free energies that then enter as position-dependent penalties in the computation of the interaction energies, preserving `RNAplex` $\mathcal{O}(m\cdot n)$ run time. The main advantage is that the accessibility profiles can be precomputed and stored, making this approach particularly attractive for large-scale screening studies. In addition, we extended `RNAplex` so that it can also handle multiple alignment. This inclusion of comparative information into the target prediction process leads to a substantial increase in specificity.

2 METHODS

2.1 `RNAplex` novelties

The extension of `RNAplex` brings two novelties that increase its specificity. First, we introduce position-specific per-nucleotide penalties that approximate the effects of the competition between intra- and intermolecular interactions. Second, `RNAplex` is now able to compute the interactions between two alignments, allowing `RNAplex` to favor evolutionary conserved interactions. Similar to the single sequence version, the multiple sequences alignment version can also consider the accessibility of the targets.

2.2 Approximate opening energies

We first outline the design of `RNAplex`, which employs a two-steps approach. In the first step, the scanning phase, `RNAplex` identifies positions where putative interactions may end. For small interior loops (1×1, 2×1 and 2×2), as well as bulges of size 1, `RNAplex` still employs the original look-up tables provided by the Turner Energy Model. For larger interior loops and bulges, however, `RNAplex` uses a linear approximation of the size dependence of loop energies (Tafer and Hofacker, 2008). The resulting energy model is exact for small loops and slightly overestimates the loop energies of large interior, bulge loops and strongly asymmetric loops. A further advantage of the linear energy model is that `RNAplex` needs to store only the last four columns of the dynamic programming matrix during the scan phase. Once all high-scoring interactions are localized along the target sequence, `RNAplex` uses the standard energy model to recompute the energy and structure of the putative hybrids.

During the scan phase, in order to extend a hybrid by one nucleotide, we need to know the cost of freeing this nucleotide from all the intramolecular interactions it might be involved in. In thermodynamic equilibrium, this energy cost can be derived from the probability that the interacting stretch of nucleotides is unpaired. Since it is too expensive to compute this for all intervals, we seek a step-wise procedure. Consider an intermediary hybrid structure $\mathcal{S}^x_y$ between two sequences x and y that starts at base pair (x_i,y_j) and spans w_x nucleotides of sequence x and w_y nucleotides of sequence y. We need to determine the *conditional* probability ${}^{w_x}\mathrm{P}^x_u[i+w_x]$ that nucleotide x_{i+w_x} is not involved in any intramolecular interaction, *given* that its predecessors $i+w_x-1$ is unpaired, and the analogous quantity ${}^{w_y}\mathrm{P}^y_u[j-w_y]$. The subscript u emphasizes that the nucleotides x and y are supposed to be unpaired. Note that this is not the same as the problem of assessing the probability $P_u[i+w_x]$ that the individual nucleotides x_{i+w_x} is unpaired, because base pairing probabilities of adjacent nucleotides are highly correlated (Bompfüenewerer *et al.*, 2008b).

The desired conditional probability can be written as:

$$ {}^{w_x}\mathrm{P}^x_u[i+w_x]=P^x_u([i+w_x]\,|\,[i,i+w_x-1]), \tag{1}$$

where the notation means that the interval $[i,i+w_x-1]$ is unpaired. An analogous expression holds for sequence y. Using the definition of the

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

conditional probability, we can write:

$$\begin{aligned}{}^{w_x}\mathrm{P}^x_u[i+w_x]&=\frac{P^x_u([i,i+w_x-1]\cup[i+w_x])}{P^x_u[i,i+w_x-1]}=\\&=\frac{P^x_u[i,i+w_x]}{P^x_u[i,i+w_x-1]}\end{aligned}\tag{2}$$

Equation (2) tells us that the conditional probability ${}^{w_x}\mathrm{P}^x_u[i+w_x]$ depends only on the probabilities $P^x_u[i,i+w_x]$ and $P^x_u[i,i+w_x-1]$ that the corresponding intervals are unpaired. Conversely, the probability that an intervals is unpaired can be computed from the conditional probabilities and the probabilities that individual nucleotides are unpaired:

$$P^x_u[i,i+w_x]=P^x_u[i]\cdot\prod_{j=1}^{w_x}{}^{j}\mathrm{P}^x_u[i+j]\tag{3}$$

A closer look at Equation (2) shows that the exact start position of the hybrid $\mathcal{S}^x_y$ has to be known in order to compute the desired conditional probability. Since `RNAplex` stores only a small number (four) of columns of the dynamic programming matrix, this cannot be done exactly. Instead we employ the approximation

$$\begin{aligned}\frac{P^x_u[i,i+w_x]}{P^x_u[i,i+w_x-1]}&\approx\frac{P^x_u[i+w_x-\delta+1,i+w_x]}{P^x_u[i+w_x-\delta+1,i+w_x-1]}\\&={}^{\delta}\overline{\mathrm{P}}^x_u[i+w_x]\end{aligned}\tag{4}$$

where δ represents the number of nucleotides considered prior to nucleotide x_{i+w_x} and ${}^{\delta}\overline{\mathrm{P}}^x_u[i+w_x]$ represents the conditional probability that x_{i+w_x} is unpaired for a given δ. This approximation is exact for $\delta=w_x$ and becomes worse with decreasing δ. This is a direct consequence of the fact that the state of the nucleotides in the interval $[i,i+w_x-\delta+1]$ is not taken into account for the computation of the conditional probability of nucleotide x_{i+w_x}.

Equation (3) can now be rewritten in the form

$$\begin{aligned}&P^x_u[i,i+w_x]\approx\\&{}^{\delta}\mathrm{P}^x_u[i,i+w_x]=P^x_u[i,i+\delta-1]\cdot\prod_{j=\delta}^{w_x}{}^{\delta}\overline{\mathrm{P}}^x_u[i+j]\end{aligned}\tag{5}$$

The probability $P^x_u[i,i+w_x]$ of being unpaired is related to a corresponding opening energy

$$\Delta G^x_u[i,i+w_x]=-\mathrm{RT}\ln P^x_u[i,i+w_x].\tag{6}$$

The energy cost of adding one nucleotide to the hybrid therefore can be written as

$$\begin{aligned}&\Delta{}^{\delta}\overline{\mathrm{G}}^x_u[i+w_x]=-\mathrm{RT}\ln{}^{\delta}\overline{\mathrm{P}}^x_u[i+w_x]=\\&\Delta G^x_u[i+w_x-\delta+1,i+w_x]\\&-\Delta G^x_u[i+w_x-\delta+1,i+w_x-1].\end{aligned}\tag{7}$$

The opening energy of a region of size of w thus is given by

$$\begin{aligned}&\Delta{}^{\delta}\mathrm{G}^x_u[i,i+w_x]=-\mathrm{RT}\ln{}^{\delta}\mathrm{P}^x_u[i,i+w_x]=\\&\Delta G^x_u[i,i+\delta-1]+\sum_{j=\delta}^{w_x}\Delta{}^{\delta}\overline{\mathrm{G}}^x_u[i+j].\end{aligned}\tag{8}$$

Since `RNAplex` only stores the current four columns of the recursion matrix, we set $\delta=4$ in practice.

2.3 Modified recursions of **`RNAplex`**

The energy $\Delta{}^{4}\overline{\mathrm{G}}^x_u[i]$ of freeing nucleotide x_i from all its intramolecular interactions can now easily be integrated into the dynamic programming recursion of `RNAplex`.

Let $C_{i,j}$ be the best interaction energy between the subsequences $x_1\dots x_i$ and $y_j\dots y_m$. Similarly, $B^x_{i,j}$ and $B^y_{i,j}$ store the optimal interactions energy given that residue x_i or residue y_j, respectively, is part of a bulge; $I_{i,j}$ stores the optimal interaction energy given that x_i and y_j are in an interior loop.

The asymmetry penalty A models asymmetric extension of interior loops. $\mathcal{S}(i,j,i-1,j+1)$ represents the energy gained by stacking base pair (x_i,y_j) onto (x_{i+1},y_{j-1}). $\mathcal{M}(i,j,i-1,j+1)$ represents the mismatch energy of the unpaired nucleotide (x_{i-1},y_{j+1}) adjacent to the pair (x_i,y_j). The energy contribution of the small interior loops is represented by $\mathcal{I}$. Furthermore, we use the following abbreviations for the opening energies: $d^x_1=\Delta{}^{4}\overline{\mathrm{G}}^x_u[i]$, $d^x_2=d^x_1+\Delta{}^{4}\overline{\mathrm{G}}^x_u[i-1]$, $d^x_3=d^x_2+\Delta{}^{4}\overline{\mathrm{G}}^x_u[i-2]$; and $d^y_1=\Delta{}^{4}\overline{\mathrm{G}}^y_u[j]$, $d^y_2=d^y_1+\Delta{}^{4}\overline{\mathrm{G}}^y_u[j+1]$, $d^y_3=d^y_2+\Delta{}^{4}\overline{\mathrm{G}}^y_u[j+2]$. The full dynamic programming recursion then reads

$$C_{i,j}=\min\begin{cases}C_{i-1,j+1}+\mathcal{S}(i,j;i-1,j+1)+d^x_1+d^y_1\\C_{i-1,j+2}+\mathcal{S}(i,j;i-1,j+2)+P_{\mathrm{bulge}}+d^x_1+d^y_2\\C_{i-2,j+1}+\mathcal{S}(i,j;i-2,j+1)+P_{\mathrm{bulge}}+d^x_2+d^y_1\\C_{i-2,j+2}+\mathcal{I}(i,j;i-2,j+2)+d^x_2+d^y_2\\C_{i-3,j+2}+\mathcal{I}(i,j;i-3,j+2)+d^x_3+d^y_2\\C_{i-2,j+3}+\mathcal{I}(i,j;i-2,j+3)+d^x_2+d^y_3\\C_{i-3,j+3}+\mathcal{I}(i,j;i-3,j+3)+d^x_3+d^y_3\\I_{i-1,j+1}+\mathcal{M}(i,j;i-1,j+1)+d^x_1+d^y_1\\B^x_{i-1,j+1}+d^x_1\\B^y_{i-1,j+1}+d^y_1\end{cases}$$

$$I_{i,j}=\min\begin{cases}C_{i-1,j+1}+\mathcal{M}(i-1,j+1;i,j)+\\\qquad+g^I_{\mathrm{open}}+2g^I_{\mathrm{ext}}+d^x_1+d^y_1\\I_{i-1,j}+g^I_{\mathrm{ext}}+A+d^x_1\\I_{i-1,j+1}+2g^I_{\mathrm{ext}}+d^x_1+d^y_1\\I_{i,j+1}+g^I_{\mathrm{ext}}+A+d^y_1\end{cases}$$

$$B^x_{i,j}=\min\begin{cases}C_{i-1,j}+g^B_{\mathrm{open}}+g^B_{\mathrm{ext}}+d^x_1\\B^x_{i-1,j}+g^B_{\mathrm{ext}}+d^x_1\end{cases}$$

$$B^y_{i,j}=\min\begin{cases}C_{i,j+1}+g^B_{\mathrm{open}}+g^B_{\mathrm{ext}}+d^y_1\\B^y_{i,j+1}+g^B_{\mathrm{ext}}+d^y_1\end{cases}$$

2.4 Hybrid structure and hybrid energy

The computation of the hybrid structure and interaction energy follows the strategy of `RNAup`. We assume that the binding region may contain mismatches and bulge loops. Thus, the most stable interaction between two segments (x_i,y_j) and (x_k,y_l) is obtained by minimizing over all possible interior loop closed by (x_p,y_q)

$$\begin{aligned}C(x_i,y_j,x_k,y_l)=&\min_{\substack{x_k<x_p<x_i\\y_l>y_q>y_j}}C(x_i,y_j,x_p,y_q)+\\&I(x_p,y_q,x_k,y_l)+\Delta G^x_u[i,k]+\Delta G^y_u[j,l]\end{aligned}\tag{9}$$

The overall most stable interaction is then obtained by minimizing over both duplex closing pairs (x_i,y_j) and (x_k,y_l):

$$E_{\min}=\min_{\substack{x_1<x_k<x_i<x_n\\y_1<y_j<y_l<y_m}}C(x_i,y_j,x_k,y_l)\tag{10}$$

where n and m are the length of sequences x and y, respectively. This leads to a theoretical run time of $\mathcal{O}(n^3\cdot m^3)$ and a memory footprint of $\mathcal{O}(n^2\cdot m^2)$.

Here we should note that one end of the hybrid, namely the base pair (x_i,y_j), was already found in the scanning phase of `RNAplex`. As a consequence, we only need to minimize over one closing pair instead of two. Equation (10) can thus be rewritten as:

$$E_{\min}=\min_{\substack{x_1<x_k<x_i\\y_j<y_l<y_m}}C(x_i,y_j,x_k,y_l)\tag{11}$$

Equations (10) and (11) show that the knowledge of base pair (x_i,y_j) allows to reduce memory and run time by a factor $n\cdot m$. Furthermore, the size of

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

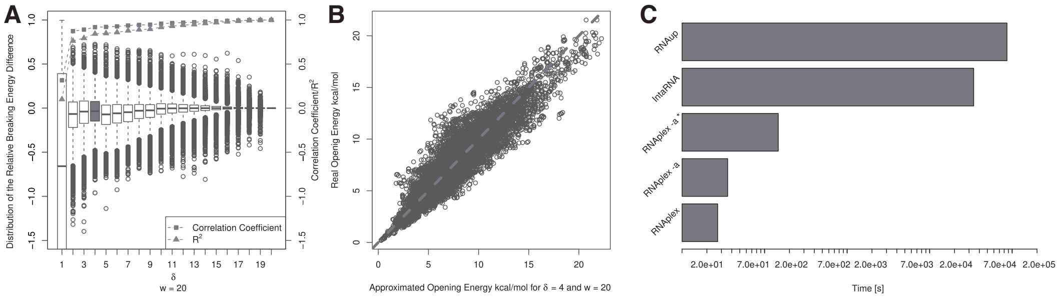


Fig. 2. (**A**) Boxplot representation of the distribution of the relative opening energy between our model and the standard energy model for different δ and a fixed target size of 20 nt. As expected, larger δ lead to smaller discrepancies. `RNAplex` uses $\delta = 4$. At this level of approximation, the Pearson's correlation coefficient between the approximated model and the real model reaches 0.92. (**B**) Scatterplot of the standard opening energies for 114 60 target sites of size 20 against the approximated opening energies as computed by `RNAplex`. (**C**) Bar plots representing the time necessary to complete the target search for 19 bacterial sRNAs in 100 random sequences of length 1200 nt for different RNA–RNA interaction tools. `RNAplex` -c, i.e. the old version of `RNAplex` is the fastest application with a completion time of 27 s. `RNAplex` -a, i.e. the new version of `RNAplex` considering accessibility, needs 36 s to achieve the same task. This grows to 120 s if one considers the time necessary to compute the accessibility profile. `RNAplex` -a is 1000 times faster than `IntaRNA` (Busch *et al.*, 2008) and 2422 times faster than `RNAup` (Mückstein *et al.*, 2008).

the interaction regions as well as the size of interior loops can be limited to arbitrary lengths ω and L, respectively, leading to a run time of $\mathcal{O}(\omega^2 \cdot L^2)$ and a memory usage of $\mathcal{O}(\omega^2)$, that is, the same complexity as `RNAduplex` or `RNAhybrid`.

2.5 Accuracy

We evaluated the performance of `RNAplex` at two levels. First, we looked at how well the opening energy derived by `RNAplex` from `RNAup` profiles matched the original `RNAup` values. Within the model of RNA secondary structures, this assess the quality of the approximations outlined in the previous section compared with the exact unpairing energies. Note that a comparison with experimentally measured opening energies is not possible since such measurements do not appear to be available in the published literature. The second test surveys how well `RNAplex` recovers the boundaries of known duplexes. This evaluates how well the different approximations made in `RNAplex` influence the quality of the predictions. The knowledge of the exact localization of RNA–RNA interactions is important, because ncRNAs may regulate their targets in different ways depending on the location of the binding sites.

In order to investigate the accuracy of the accessibility profiles, we used a set of 11 460 randomly generated sequences of length 400 nt for which the accessibility profiles was computed with `RNAup`. For each sequence, we then determined the difference of the `RNAup` opening energy and the `RNAplex` opening energy for the region located between nucleotides 181 and 200. Figure 2 shows the relative energy differences between both models as bar plots for different values of δ. The largest variations are seen for $\delta = 1$ with differences larger than 100%. R^2 (triangle) and the Pearson's correlation coefficient (square) reach their minimum there (0.09 and 0.37, respectively). Both coefficients then steadily improve with δ and reach their theoretical maximum of 1 for $\delta = w$. For $\delta < w$, our approximation slightly overestimates the opening energy. This can be seen for $\delta = 4$, the value used in `RNAplex` in the scatterplot in the middle of Figure 2. Half of the relative deviation are contained between +7% and −14%.

The accuracy of the energy model (interaction and opening energy) used in `RNAplex` was compared with that of `RNAup`, `biRNA` (Chitsaz *et al.*, 2009), and the old version of `RNAplex` (`RNAplex -c`) on a dataset of 17 known bacterial small RNA–mRNA interactions (Chitsaz *et al.*, 2009) (see Supplementary Material). In this dataset, both the opening energy of the interacting sequences and the hybridization energy affects the prediction.

`RNAplex -c` (old version) missed four interactions, while all `RNAplex -a` (with accessibility information) predictions overlapped with the corresponding experimentally determined interactions, as did the predictions of `RNAup` and `biRNA` (see Supplementary Table S2). These results emphasize the importance of accessibility for the correct prediction of RNA–RNA interactions. Furthermore, it confirms that the approximations used in `RNAplex` are sufficient to reach a level of accuracy similar to that of `RNAup` and `biRNA`.

The location of the predicted closing pairs was compared to the confirmed locations. For each prediction tool, the average over all 17 interactions of the sum of the magnitude of the deviation between the predicted and confirmed locations of the four closing nucleotides was computed. All three accessibility-based methods performed similarly with an average deviation of 16.76 for `RNAup`, 19.88 for `biRNA` and 20.60 for `RNAplex -a`, much smaller than the average deviation of `RNAplex -c` (59.76 nt) (see Supplementary Table S2).

It should be noted that `RNAup` and `RNAplex`, in contrast to `biRNA`, cannot handle interactions involving two or more interacting regions, such as the two kissing-hairpin complexes found in *OxyS-fhlA*. Still, in contrast to `RNAup`, `RNAplex` can return suboptimal predictions, without run time overhead, that can be used to identify disjoint interaction regions. For *OxyS-fhlA*, the confirmed binding regions are located at positions [22,30] and [98,104] on *OxyS* and [87,95] and [39,45] on *fhlA*, in accord with the two best suboptimals returned by `RNAplex` which are located on [23,28] and [96,100] on *OxyS* and [87,92] and [41,45] on *fhlA*.

2.6 Computational efficiency

The run time of the new version of `RNAplex` was compared with that of the old version (`RNAplex -c`, no accessibility), `RNAup` and `IntaRNA` (Busch *et al.*, 2008) on a dataset containing 19 *`E.coli`* sRNAs and 100 *`E.coli`* mRNAs (see Supplementary Material). For each gene, we defined the putative target region as the sequence interval from 200 nt upstream and 1000 nt downstream of the start codon.

`RNAplex` completed this task in 36 s, while `IntaRNA` and `RNAup` needed 34150 and 86487 s, respectively. The run time of `RNAplex` thus is reduced by a factor of 2400 and 950 compared with `RNAup` and `IntaRNA`, respectively. If we count the time needed to compute the accessibilities needed by `RNAplex`, the total run time reaches 120 s, still more than two orders of magnitude less than the other tools (Fig. 2).

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

We further compared the run time and the memory consumption of `RNAup` and `IntaRNA` against that of the new `RNAplex`, by generating a set of random target sequences of size 400, 800, 1600, 3200 and 6400 nt and query sequences of size 100, 200, 400 and 800 nt and searching for targets with all three tools. On this dataset, the new `RNAplex` is between 575 and 1600 times faster than `IntaRNA` and between 1500 and 65 400 times faster than `RNAup`. The memory consumption is also drastically reduced. `RNAplex` needs at least 17 and at most 1330 times less memory than `IntaRNA` , and 15–626 times less memory than `RNAup` (see Supplementary Table S1). Compared to the old version without accessibilities, the new `RNAplex` needs only four times more memory.

2.7 Conserved interactions

The absence of conserved target site in closely related species may indicate that the proposed interaction does not occur in nature. The presence of compensatory mutations between the sRNA and the target site, on the other hand, can lend further credibility to single sequence target predictions (Chen *et al.*, 2007). Alignments thus can improve the specificity of target search by focusing on evolutionary conserved interactions.

We, therefore, extended `RNAplex` to alignments. The approach follows the same idea as `RNAalifold` (Bernhart *et al.*, 2008; Hofacker *et al.*, 2002), where a thermodynamic energy minimization folding algorithm is coupled with a simple scoring model to assess structural evolutionary conservation. Base pairs are, therefore, restricted to pairs of positions in the alignments in which most or all sequences can form canonical pairs.

The evolutionary model used in `RNAplex`, while straightforward, performs well in predicting consensus secondary structure. Its simplicity allows it to be integrated into `RNAplex` without run time overhead (see Supplementary Material).

A potential weakness is the `RNAalifold` scoring model, which was trained and optimized for intramolecular interaction, instead for the intermolecular interactions to which it is applied here. More complex scoring schemes such as the one used in `PETfold` and `PETcofold`, where a maximum expected scoring approach combines the evolutionary probabilities of a consensus structure given an alignment with the thermodynamic probabilities of the associated structures in each sequence (Seemann *et al.*, 2008, 2010, 2011), perform slightly better than the `RNAalifold` scoring scheme. However, they can be incorporated only at the cost of a greatly increased run time, and thus are incompatible with the purpose of `RNAplex`.

Similar to the single sequence version, the alignment version of `RNAplex` only allows interior loops in the RNA–RNA hybrids. Like the single sequence, accessibility can be taken into account by averaging the position-dependent extension costs computed for the individual sequences in the alignment (see Supplementary Materials for a full description of the recursion).

2.8 Datasets

A complete description of all datasets used in this study can be found in the Supplementary Materials.

3 APPLICATION

As an application example, we consider the genome-wide prediction of sRNA targets in `E.coli`. As a reference set, we use the experimentally confirmed interactions published by Urban *et al.* (2007). We expect that, for a given sRNA, the number of predicted interactions with other (false positive) targets should decrease when accessibility of the target mRNA in included. Ideally, it should reach the low levels observed for `RNAup` (Mückstein *et al.*, 2008).

For each sRNA, the amount of false positives was estimated by counting genome wide the number of sRNA-target interactions that are more stable than the experimentally reported sRNA-target duplex. For each 4463 `E.coli` genes, a mRNA of length 1200 nt, including 200 nt upstream and 1000 nt downstream of the start codon were defined. Accessibility profiles were computed with `RNAplfold`, with a folding windows (option `-W`) of 240 nt and a maximal base pair distance of 160 (option `-L`). An interaction was reported if the corresponding sRNA–mRNA interaction energy is smaller than the experimentally confirmed interaction, and if it occurs in region encompassing 80 nt, 50 nt upstream and 30 nt downstream of the start codon.

The inclusion of the accessibility profiles in the new version of `RNAplex` leads to a substantial improvement as can be seen from Table 1. All native interaction sites are among the predictions, and the detailed target site localization is improved. Most importantly, the number of predictions with better interaction energies, i.e. the false positives, is reduced to a level similar to that of `RNAup`.

In order to better assess the number of false positives, the same method was applied on the dinucleotide-shuffled sRNAs and mRNAs. To this end, we compared the interaction energy of the non-shuffled, experimentally confirmed interactions, to the energy distribution of the shuffled sequences. Interestingly, in seven out of nine cases, the number of false positives is smaller (see Supplementary Material) in the shuffled case than in the non-shuffled one. This can be explained by the fact that in various bacteria, the region around the ribosomal entry site, which is also the preferred region of sRNA binding, is more accessible than the rest of the mRNA (see Supplementary Material). This in turn implies that compared with shuffled sequences, sRNAs have a greater chance to bind to the region around the start codon in non-shuffled mRNAs. Depending on the ncRNAs, one can expect between 7.5×10^{-7} false positives per nucleotide for micC and $1.5 \times 10-4$ false positives for gcvB (see Supplementary Material).

3.1 Multiple alignment

While `RNAplex` recovers all interactions, some of them like `RyhB-sodB` or `GcvB-oppA` are ranked lowly. A comparative version of `RNAplex` was designed (see Section 2) to reduce the number of false positives. Similar to consensus RNA folding, the quality of the input alignments is crucial to obtain meaningful results (Bernhart *et al.*, 2008).

The comparison of the performance of the single sequence with the comparative version of `RNAplex` was achieved by generating multiple sequences alignments `clustalw` (Larkin *et al.*, 2007) for the eight sRNAs from Table 1 and with `MUSCLE` (Edgar, 2004) for the 4463 `E.coli` mRNAs. The list of bacteria used for the alignment are found in the Supplementary Material.

In many cases, `MUSCLE` and `clustalw` were not able to satisfactorily align the sequences. This was caused e.g. by misannotations of the start codon as for the `ompA` gene in `E.coli APEC 01`, which was incorrectly annotated 70 nt upstream of the true start codon. In order to better handle these cases, we devised a method to produce multiple alignments of highly similar and strongly binding target sites (see Supplementary Materials).

Because highly conserved interactions are more credible than non-conserved interactions, ranking of interactions based on multiple sequences alignments should not only take the interaction energy into account, but also the number of organisms (in which a predicted interactions is detectable). This can be achieved by using Z-scores

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

Table 1. Summary of the predicted binding sites for the nine functional interactions reported by Urban *et al.* (2007)

sRNA	mRNA	Pos.lit.	Pos$_{\texttt{RNAplex}}$	ΔG `RNAup`	ΔG `RNAplex`	ΔG `RNAplex -A`	ΔG	Z-score	Z-score N^{o}seq
RyhB	sodB	−7, +5	−4, +5	−10.50 (60)	−11.08 (50/87)	−9.31 (12)	65	57	2 (7)
DsrA	hns	+6, +21	+7, +19	−10.90 (17)	−12.74 (2/128)	−11.25 (10)	1	12	0 (0)
MicA	ompA	−21, −6	−21, −6	−13.46 (0)	−14.35 (1/67)	−14.04 (14)	0	11	0 (0)
MicC	ompC	−30, −15	−30, −15	−15.85 (1)	−16.24 (2/97)	−17.50 (9)	0	0	0 (0)
MicF	ompF	−8, +10	−16, +10	−17.00 (3)	−13.65 (8/34)	−18.28 (6)	0	0	0 (2)
Spot42	galK	−19, +14	−19, +21	−18.92 (0)	−13.02 (25/38)	−7.31 (9)	25	28	5 (12)
SgrS	ptsG	−28, −8	−28, +4	−17.17 (1)	−17.53 (0/170)	−11.17 (10)	5	4	0 (1)
GcvB	dppA	−31, −10	−31, −14	−16.90 (16)	−17.11 (8/80)	−13.15 (9)	14	14	7 (19)
GcvB	oppA	−4, 21	−8, 16	−11.64 (58)	−12.00 (36/263)	−14.43 (5)	27	26	14 (19)

The first and second columns show the name of interaction partners. Columns 3 and 4 give the predicted and experimentally reported binding regions, respectively. Columns 5 and 6 report the binding ΔG computed by `RNAup` and `RNAplex`, respectively. The numbers in parenthesis in the sixth column represent the number of interactions, located within a window of 80 nt centered around the start codon, with a lower interaction energy than the experimentally reported interaction for the predictions made by `RNAplex` with and without considering the opening energy, respectively. Column 7 gives the interaction energy for the multiple sequences interactions. The numbers in parenthesis in column 7 represent the number of sequences in the final alignments. Column 8 shows the rank of the interaction when looking only at the interaction energy. Column 9 shows the rank of the interactions based on the Z-score corrected for the number of sequences in the alignment. Finally, column 10 shows the rank of the interaction based on the Z-score, given that only interactions with a greater or equal number of sequences in the alignment are taken into account. The number in parenthesis in the last column represent the number of better scoring elements in the case of alignment when no accessibility information are taken into account.

as alternative ranking criterion. The Z-scores can be computed for all interactions having the same number of sequences in the alignments. This is important as highly conserved interactions tend to have a higher consensus interaction energy than interactions that are conserved in only few organisms (see Supplementary Figure S2).

In this way, extremely stable interactions can be compared without having to worry about the number of sequences in the alignments. The main drawback of this method is that highly conserved interactions with more than 10 sequences are rare, making the Z-score analysis unreliable. This is the case, for example, for the *micA-ompA* pair, which has the highest interaction energy among the interactions involving 14 species. In this case, the rank of MicA drops from 2 for the single sequence approach to 11 for the alignment approach.

Table 1 shows that the rank based on the interaction energy or the Z-score is similar to that of the single sequence energy ranking. However, when considering only interactions having a greater or equal number of sequences and a higher Z-score (column 10), the number of interactions that score better than the native one in the single sequence case (column 6) decreases significantly, with the greatest reduction being seen for *ryhB*. This is especially interesting because the *ryhB-sodB* is difficult to predict, probably due to its dependence upon *Hfq*, a protein known to facilitate sRNA–mRNA duplex formation (Sittka *et al.*, 2007). Similar to the single sequence case, the use of accessibility information in the case of multiple sequences alignments allows to improve the rank of the known interactions. This can be seen in the last column of Table 1.

It should be noted that some false positives turned out to be real interactions: for example, *iscS* and *acnB* score better than *sodB* as targets for *ryhB* and are true targets (Desnoyers *et al.*, 2009; Massé and Gottesman, 2002). Similar trends can be seen if the Z-score threshold is set to 0 and the number of sequences in the multiple alignment remains unchanged. If we look at the gene ontology of these targets in the case of *ryhB* (43 targets), we see that 35 are involved in catalytic activities ($P = 0.006$), 9 are involved in iron–sulfur cluster binding ($P = 0.007$), 39 are involved in binding ($P = 0.01$). *ryhB* targets are also significantly overrepresented in the CO_2 fixation ($P = 0.0001$) as well as citrate cycle cellular pathways ($P = 0.0002$), in line with the gene ontology analysis. More examples can be found in the Supplementary Materials.

4 DISCUSSION

We presented a new version of `RNAplex`, a tool designed to rapidly and reliably predict RNA–RNA interactions. Compared with the previously published version, `RNAplex` now considers target site accessibility, by using accessibility profiles generated by `RNAplfold` to approximate the energy of removing a nucleotide from all intramolecular interactions. The introduction of position-specific, structure-dependent extension cost allows to greatly improve the specificity of `RNAplex`, bringing it close to that of `RNAup`, without modifying the linear run time of the original `RNAplex`.

Clearly, the main feature of `RNAplex` is its run time efficiency. On a dataset of 19 ncRNAs and 100 target mRNAs on length 1200, `RNAplex` runs 2400 faster than `RNAup` without noticeably loss of specificity, thus making ncRNAs target searches more affordable. In its present implementation, `RNAplex` can be used not only to predict ncRNA targets in small genomes, but can also be used to find miRNA targets and siRNA off-targets in large mammalian genomes and transcriptomes and it can be applied to microarray probes design. In contrast to `RNAup` or `RNAhybrid`, `RNAplex` can return suboptimal solutions efficiently on the fly without the need of recomputing the full recursion matrix.

The ability of `RNAplex` to perform comparative target search allows to discard poorly conserved interaction and to lend further credibility to interactions showing compensatory mutations. Based on a dataset of experimentally confirmed interactions, we show that `RNAplex` in its present form is an useful tool to predict new sRNA targets. We further show that suboptimal predictions from `RNAplex` may actually be real targets. Application of the comparative version

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

of RNAplex on larger genomes and other ncRNAs, e.g. miRNAs, is straightforward.

In order to make RNAplex more usable for the community, we plan to set up a web server especially designed to predict targets for sRNAs in bacteria. We further plan to use RNAplex to better understand the regulatory circuits found in E.coli (Shimoni *et al.*, 2007). Finally, a probe design method based on RNAplex is currently being developed.

Funding: European Union under the auspices of the FP-7 QUANTOMICS project (HT,PSD, in part); DFG priority program SPP1258 Sensory and Regulatory RNAs in prokaryotes (HT,PSD, in part); and Austrian GENAU project 'Regulatory Noncoding RNA' [ILH,FA].

Conflict of Interest: none declared.

REFERENCES

Alkan,C. *et al.* (2006) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **13**, 267–282.

Andronescu,M. *et al.* (2003) RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.

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.

Bachellerie,J. *et al.* (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Backofen,R. *et al.*; The Athanasius F. Bompfünewerer RNA Consortium (2007). RNAs everywhere: genome-wide annotation of structured RNAs. *J. Exp. Zool. B Mol. Dev. Evol.*, **308B**, 1–25.

Banerjee,D. and Slack,F. (2002) Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays*, **24**, 119–129.

Benne,R. (1992) RNA editing in trypanosomes. the us(e) of guide RNAs. *Mol. Biol. Rep.*, **16**, 217–227.

Bernhart,S. *et al.* (2006a) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.

Bernhart,S. *et al.* (2006b) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.

Bernhart,S. *et al.* (2008) Rnaalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Bompfünewerer,A. *et al.* (2008) Variations on RNA folding and alignment: lessons from benasque. *J. Math. Biol.*, **56**, 129–144.

Busch,A. *et al.* (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.

Chen,C. *et al*, (2007) Exploration of pairing constraints identifies a 9 base-pair core within box c/d snoRNA-rRNA duplexes. *J. Mol. Biol.*, **369**, 771–783.

Chitsaz,H. *et al.* (2009) *Algorithms in Bioinformatics*, Vol. 5724 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Desnoyers,G. *et al.* (2009) Small RNA-induced differential degradation of the polycistronic mRNA iscrsua. *EMBO J.*, **28**, 1551–1561.

Dimitrov,R. and Zuker,M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, **87**, 215–226.

Dirks,R. *et al.* (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**, 65–88.

Edgar,R. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Eguchi,Y. and Tomizawa,J. (1990) Complex formed by complementary RNA stem-loops and its stabilization by a protein: function of coie1 rom protein. *Cell*, **60**, 199–209.

Fire,A. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.

Gardner,P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, 136–140.

Hofacker,I. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Hofacker,I. *et al.* (2002) Secondary structure prediction for aligned rna sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Huang,F. *et al.* (2010) Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics* , **26**, 175–181.

Kugel,J. and Goodrich,J. (2007) An RNA transcriptional regulator templates its own regulatory RNA. *Nat. Chem. Biol.*, **3**, 89–90.

Larkin,M. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Massé,E. and Gottesman,S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl Acad. Sci. USA*, **99**, 4620–4625.

Mückstein,U. *et al.* (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.

Mückstein,U. *et al.* (2008) Translational Control by RNA-RNA Interaction: Improved Computation of RNA-RNA Binding Thermodynamics. In Elloumi,M. *et al.* (eds) *Bioinformatics Research and Development*, Vol. 13, Springer, Berlin/Heidelberg, pp. 114–127.

Pervouchine,D. (2004) Iris: intermolecular rna interaction search. *Genome Inform.*, **15**, 92–101.

Rehmsmeier,M. *et al.* (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Seemann,S. *et al.* (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.

Seemann,S. *et al.* (2010) Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions. *Algorithms Mol. Biol.*, **5**, 22.

Seemann,S.E. *et al.* (2011) PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, **27**, 211–219.

Shimoni,Y. *et al.* (2007) Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol. Syst. Biol.*, **3**, 138.

Sittka,A. *et al.* (2007) The RNA chaperone hfq is essential for the virulence of Salmonella typhirium. *Mol. Microbiol.*, **63**, 193–217.

Stark,A. *et al.* (2003) Identification of Drosophila MicroRNA targets. *PLoS Biol*., **1**, e60.

Tafer,H. and Hofacker,I. (2008) Rnaplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Urban,J. *et al.* (2007) A conserved small RNA promotes discoordinate expression of the glmUS operon mRNA to activate GlmS synthesis. *J. Mol. Biol.*, **373**, 521–528.

Zorio,D. *et al.* (1997) Cloning of caenorhabditis u2af65: an alternatively spliced RNA containing a novel exon. *Mol. Cell Biol.*, **17**, 946–953.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Institute for Theoretical Chemistry and Structural Biology on April 26, 2013

9 sRNA binding and its influence on translation initiation

9.1 Statement of personal contribution

ILH, CF, FA designed the study. FA implemented and tested the program, FA wrote the paper.

9.2 Article

Fabian Amman, Christoph Flamm, and Ivo L. Hofacker.
"Modelling Translation Initiation under the Influence of sRNA."
International journal of molecular sciences 13, no. 12 (2012): 16223-16240.

Received on October 21, 2012,
Revised on November 21, 2012,
Accepted on November 27, 2012;

© 2012 by the authors; licensee MDPI, Basel, Switzerland. Reprint with permission from MDPI, and the authors.

Int. J. Mol. Sci. **2012**, *13*, 16223-16240; doi:10.3390/ijms131216223

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

Modelling Translation Initiation under the Influence of sRNA

Fabian Amman *, Christoph Flamm and Ivo Hofacker

Institute for Theoretical Chemistry, University Vienna, Währingerstraße 17, 1090 Vienna, Austria;
E-Mails: xtof@tbi.univie.ac.at (C.F.); ivo@tbi.univie.ac.at (I.H.)

* Author to whom correspondence should be addressed; E-Mail: fabian@tbi.univie.ac.at;
Tel.: +43-1-4277-527-34; Fax: +43-1-4277-527-93.

Received: 21 October 2012; in revised form: 21 November 2012 / Accepted: 27 November 2012 / Published: 30 November 2012

Abstract: Bacterial small non-coding RNA (sRNA) plays an important role in post-transcriptional gene regulation. Although the number of annotated sRNA is steadily increasing, their functional characterization is still lagging behind. Various computational strategies for finding sRNA–mRNA interactions, and thus putative sRNA targets, were developed. Most of them suffer from a high false positive rate. Here, we present a qualitative model to simulate the effect of an sRNA on the translation initiation of a potential target. Information about the ribosome–mRNA interaction, sRNA–mRNA interaction and expression information from deep sequencing experiments is integrated to calculate the change in translation initiation complex formation, as a proxy for translational activity. This model can be used to post-evaluate predicted targets, hence condensing the list of potential targets. We show that our translation initiation model, under the influence of an sRNA, can successfully simulate thirteen out of fifteen tested sRNA–mRNA interactions in a qualitative manner. To show the gain in specificity, we applied our method to a target search for the *Escherichia coli* sRNA RyhB. Compared with simple target prediction without post-evaluation, we reduce the number of targets to less than one fourth potential targets, considerably reducing the burden of experimental validation.

Keywords: sRNA; sRNA target prediction; translation initiation

---

1. Introduction

Bacteria's competence to adapt to changing environmental conditions is one key to their ecological success. Beside the network of transcription factors, a second layer of regulation has attracted attention since 1984 when the influence of the RNA MicF on the expression of ompF was discovered [1]. Since trans-acting small non-coding RNA (sRNA) shifted into the focus of research, remarkable progress was made describing new sRNA genes in a number of bacterial species. Experimental approaches (micro-arrays, co-purification, and more recently, next generation sequencing) could successfully verify more than 80 sRNA genes in *Escherichia coli* [2]. Computational screens based on sequence conservation, structural homology or expected components, like promoters and terminators, suggest the existence of hundreds more [3]. Meanwhile the functional description of newly found sRNA genes becomes the main obstacle in broadening the existing gene regulation networks.

Functional characterization is still a challenging task. It is not clear from the outset by which mechanism an sRNA works. They bind to proteins, altering their activity [4], or they bind to target mRNA, thus influencing their stability or translation. The latter can be performed in different ways. Some sRNA block translation initiation by competing with the ribosome binding site (RBS) of the mRNA. This leads to reduced translation, which can again cause degradation of the unused mRNA molecule. A less frequent effect of a bound sRNA is to fortify the translation rate by inducing a refolding of the translation initiation region (TIR) and thus dissolving translation inhibiting structures. Additionally, some sRNA exclusively regulate only one target whereas others can interact with dozens of targets, applying a different one of the above-mentioned mechanisms each time. In contrast to miRNA in eukaryotes, where a lot of binding rules are marked out (such as a $5'$ binding seed or a preference for binding sites at the ends of $3'$ UTR [5]), the interactions of sRNA with their mRNA counterparts show a striking variability in bacteria [3].

All this complexity is reflected by the fact that there is no satisfying standalone technique to find new targets for an sRNA yet. Experimental approaches are very labor intensive, which means that they are not applicable to broad genomic screens (e.g., two-plasmid reporter gene assay [6]), or they are not suitable to properly distinguish between primary and secondary regulation effects (e.g., sRNA over-expression or deletion with downstream transcriptome profiling [7]).

Computational target prediction methods have shown to be helpful. The applied techniques range from mere sequence-based methods comparable with `Blast` [8] (e.g., TargetRNA [9]), to more sophisticated methods that calculate the hybridization energy by considering the inter-molecular base-pairing and stacking energies (implemented in, e.g., `RNAduplex`, part of the `ViennaRNA Package` [10]). The latest generation also includes intra-molecular structure, thus taking the accessibility of the putative binding site into account. This approach was implemented in `RNAup` [11], `IntaRNA` [12] and most recently `RNAplex` [13] in combination with `RNAplfold` [14,15]. The structure based tactics are similar in their attempt to find the best possible interaction or interactions between two given RNA sequences. Since any two sufficiently long sequences will show some stable interaction, the decision of which sequences to search and how the results are interpreted is up to the user. A common strategy is to concentrate on a sequence stretch of -30 nt to +20 nt around the translation start site [9], which, by reducing the search space, reduces the number of predicted nonfunctional binding sites. This strategy has proven to be quite successful since many observed interactions are indeed taking place in this region. However, some interactions are known to be further upstream. In *E. coli*, DsrA and RprA bind their target rpoS at position -94 nt and -93 nt, respectively, upstream of the translation start site where they induce an activation of translation [16]. OmrB represses csgD by binding from position -79 nt to -61 nt in front of the gene's start site [17]. Even in the reduced search space around the start codon, it seems that the thermodynamically best binding sites are not always the biologically functional ones. Some experimentally observed binding sites show an unfavorable calculated binding energy and thus are easily overseen in genome wide screens. This might be explained by the activity of chaperons such as Hfq, which stabilize the sRNA–mRNA interaction [18].

This is why we developed a new approach to extend the common binding site prediction with an automated evaluation of the functional consequences of a bound sRNA on translation initiation. This is achieved by introducing a model that simulates the initiation of translation in the system mRNA, sRNA and 16S ribosome. With this approach, it is possible to examine which of the putative interactions have the potential to interfere with translation initiation. In the following article, we will lay out how our model can simulate this influence and show that this can be helpful to evaluate predicted target sites for their biological significance.

2. Model Description

Translation initiation is the process by which components of the ribosome detect an mRNA, which leads to the assembly of the ribosomal machinery. It was demonstrated that this is the rate limiting step for translation [19]. It is triggered by the binding of the 30S ribosome unit, via the $3'$ end of the 16S ribosomal RNA, to the Shine–Dalgarno sequence (SD) and the positioning of the fMet-tRNA^{fMet} anti-codon to the correct start-codon on the mRNA. A mathematical model of this process was developed by Na and Lee [20], whose concept and nomenclature are adopted here. The model was slightly adapted and substantially expanded to include the influence of sRNA binding on translation initiation.

Kinetically, the initiation of the ribosome–mRNA interaction is driven by the energy gained from the hybridization of the 16S rRNA to the ribosome recognition site (RRS, *i.e.*, a generalization of the Shine–Dalgarno sequence) and the anti-start-codon–start-codon interaction. Further on, the accessibility of the complete ribosome docking site (RDS, *i.e.*, the stretch of the mRNA that is occupied by the translation initiation complex) is essential because during initiation the ribosome has no capability to dissolve inhibiting structures on the mRNA [21]. At this point, the sRNA can interfere with ribosome binding: Either it competes with the ribosome for binding within the RDS or it alters the accessibility of the RDS by binding close-by and inducing a refold, hence changing the mRNA accessibility for the ribosome.

We define the RRS as the energetically most favorable binding site of the anti-RRS (the $3'$ end of the 16S rRNA, in the case of *E. coli* this would be "UCACCUCCUU") upstream of the translation start site. Calculating all possible interactions and choosing the energetically most favorable one, provides the position of the RRS and the ribosome–mRNA hybridization energy ΔG_R. To account for the stabilizing effect of anti-start-codon–start-codon interaction, $-1.19\ ^{kcal}\!/_{mol}$ for *AUG*, $-0.075\ ^{kcal}\!/_{mol}$ for *GUG* and $0\ ^{kcal}\!/_{mol}$ for all other are added to ΔG_R [22].

The RDS was shown to be about 30 nt long [19], starting from the predicted RRS start. The RDS exposing probability of the free mRNA P_{EF} (*i.e.*, the probability that this 30 nt long sequence is accessible for the ribosome), or equivalently the free energy $\Delta E_F = -RT \ln P_{EF}$ needed to make the RDS accessible, is the main thermodynamic barrier in translation initiation.

Regarding the system consisting of mRNA, sRNA and ribosome, the following reactions (Equations 1–4) lead from the free unbound mRNA M_F to the ribosome bound mRNA M_R or can compete with this reactions. For simplicity, Equation 4 itself is not included in the model.

$$M_F + S_F \overset{K_S}{\rightleftharpoons} M_S \tag{1}$$

$$M_F^* + R_F \overset{K_R}{\rightleftharpoons} M_R \tag{2a}$$

$$M_F \overset{K_{EF}}{\rightleftharpoons} M_F^* \tag{2b}$$

$$M_S^* + R_F \overset{K_R}{\rightleftharpoons} M_{SR} \tag{3a}$$

$$M_S \overset{K_{ES}}{\rightleftharpoons} M_S^* \tag{3b}$$

$$M_R + S_F \overset{K_{SR}}{\rightleftharpoons} M_{SR} \tag{4}$$

Thereby, M_F is the free unbound mRNA, R_F the free ribosome, S_F the free sRNA, M_S and M_R the sRNA and the ribosome bound mRNA, respectively. M_{SR} represents the mRNA species with sRNA and ribosome bound at the same time. The superscript asterisk "*" marks the RDS exposing fraction of its kind. In the following, we will use the convention to address reaction species with uppercase letter, whereas lowercase letters are used when we refer to the concentration of the particular reaction species.

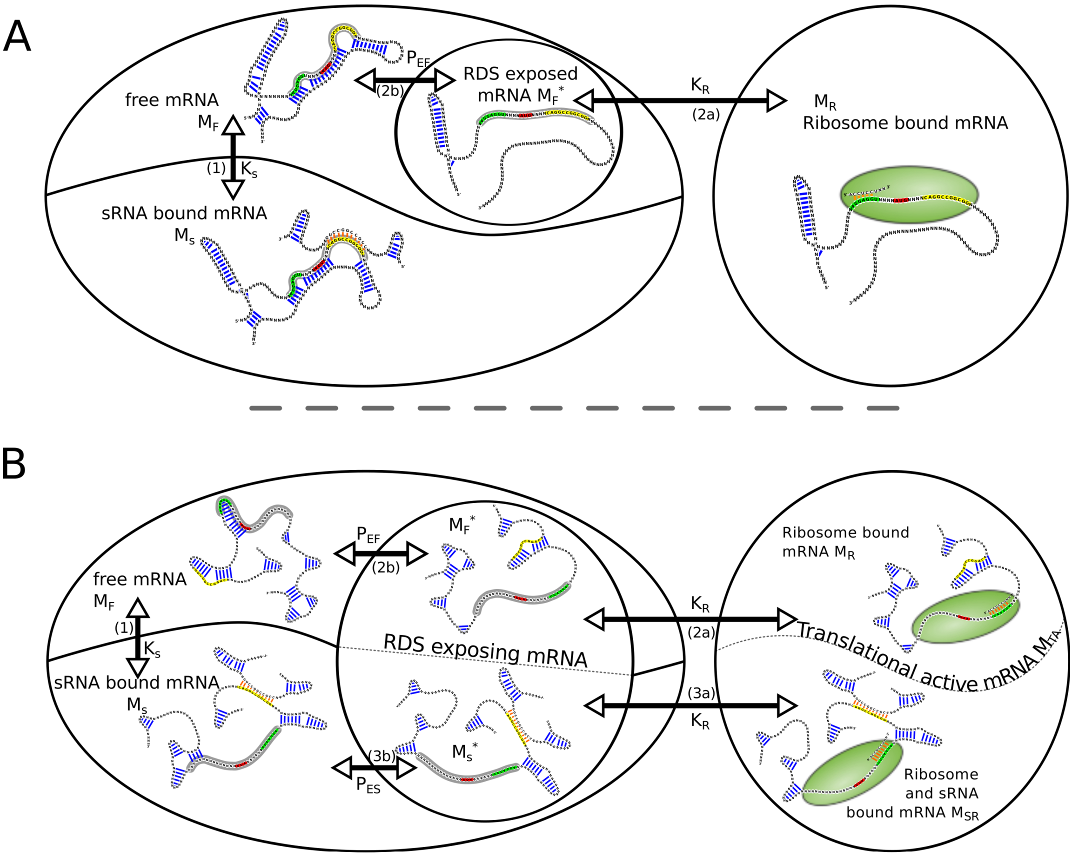
The equilibrium constants of the ribosome binding and sRNA binding reaction, $K_R = \exp(-\frac{\Delta G_R}{RT})$ and $K_S = \exp(-\frac{\Delta G_S}{RT})$, respectively, can be calculated from the free energy difference of the reaction ΔG_R and ΔG_S, where T is the temperature and R the gas constant. Please note that the reaction constant for the ribosome binding to the mRNA K_R is independent of mRNA structure, thus the same in Equation 2a,3a. The mRNA structure is already considered through the formation of M^* (Equation 2b,3b).

K_{EF} and K_{ES} denote the equilibrium constants of the unfolding reaction of the complete RDS, without and with the influence of a bound sRNA, respectively. The reaction constants are connected to the probabilities P to expose the RDS by $P = \frac{K}{1+K}$. In the following we will only work with the corresponding probabilities P_{EF} and P_{ES}.

To calculate the amount of ribosome bound mRNA, the relative positions of the sRNA binding site and the RDS have to be considered. In the case where the RDS overlap with the sRNA binding site, reaction 3a is not possible since a simultaneous binding of the ribosome and the sRNA is sterically not possible, thus species M_{SR} does not occur. If RDS and sRNA binding site are spatially separated, sRNA and ribosome can bind to the same mRNA molecule, hence two translational active mRNA species, M_R and M_{SR}, have to be considered.

The chemical reaction network above can be readily translated into a system of equations describing the equilibrium concentrations of all chemical species. In the following, we use this to calculate the amount of ribosome bound mRNA and its dependence on sRNA presence. Figure 1 depicts the different routes and reactions that lead from the unbound mRNA to translational active, namely ribosome bound mRNA.

Figure 1. Graphical illustration of all reactions and species considered in the reaction network. The RNA species are depicted with black backbones, blue intra-molecular and orange inter-molecular base-pairs. The ribosome with its anti-RRS sequence is shown as a green sphere. The RDS is highlighted in gray. The RRS, the start codon and the RNA binding site are marked with green, red and yellow, respectively. Reactions are symbolized with ↔ arrows, their corresponding equilibrium constants and a reference to the reaction equation in the main text. (**A**) In the case where the RDS and the RNA binding site overlap, two reaction branches from M_F compete with each other. One leads to sRNA bound mRNA M_S, the other leads via M_F^* to ribosome bound mRNA M_R; (**B**) In the case where the RNA binding-site and RDS are spatially separated, there are two routes from free mRNA to translationally active M_{TA}. One leads as before via M_F^* to M_R. The other route first leads to an sRNA·mRNA complex, which can further expose its RDS M_S^*, and eventually ends in the active ribosome·mRNA·sRNA complex M_{SR}.



2.1. Overlap of sRNA-BS and RDS

Since sRNA and ribosome cannot bind the same mRNA, the only translational active mRNA is the M_R species. The ribosome binds the free RDS exposing mRNA in thermodynamic equilibrium with

$$K_R\, m_F^*\, r_F = m_R \tag{5}$$

At the same time the sRNA binding competes with this reaction. sRNA binding onto free mRNA can be described with

$$K_S\, m_F\, s_F = m_S \tag{6}$$

Furthermore, the following relationships can be formulated, thereby s_F, s_T, m_F, m_S, m_R, m_T, r_F and r_T describe the concentrations of free sRNA, total sRNA, free mRNA, sRNA bound mRNA, ribosome bound mRNA, total mRNA, free ribosome and total ribosome, respectively.

$$s_F + m_S = s_T \tag{7}$$

$$m_F + m_S + m_R = m_T \tag{8}$$

$$r_F + n\, m_R = r_T \tag{9}$$

The pool of free ribosomes is depleted not only by ribosomes bound at the TIR but also by actively translating ribosomes. To account for this, we follow Na and Lee [20] and introduce the ribosome occupancy n in Equation 9. The value n is estimated from experiments on the *E. coli* lac operon that show on average 20 ribosomes bound to the mRNA [23]. Thus, each initiation event (as modelled by Equation 5) ultimately reduces the number of free ribosomes by approximately $n = 20$.

Taking this system of five equations (Equations 5–9) together with $m_F^* = P_{EF}\, m_F$ allows to compute the amount of translation initiation complex m_R as function of K_R, K_S, P_{EF}, s_T, r_T, m_T and n. In principle the variables s_F, m_F, m_F^*, r_F and m_S can be eliminated resulting in a cubic polynomial that is analytically and numerically solvable. Details can be found in the supplementary material.

2.2. No Overlap of sRNA-BS and RDS

When RDS and sRNA binding site are spatially separated, both binding sites can be occupied at the same time. As a consequence, two species in the described reaction network represent active translation initiation complexes. To contribute for this we introduce a new variable for the translational active mRNA m_{TA}.

$$m_{TA} = m_R + m_{SR} \tag{10}$$

Furthermore, we have to consider reaction 3a, describing the binding of a ribosome to an sRNA·mRNA complex

$$K_R\, m_S^*\, r_F = m_{SR} \tag{11}$$

In contrast to the first case with overlapping RDS and sRNA-BS, Equations 7–9 have to be adapted in the following way to include the new species of m_{SR}

$$s_F + m_S + m_{SR} = s_T \tag{12}$$

$$m_F + m_S + m_R + m_{SR} = m_T \tag{13}$$

$$r_F + n\,(m_R + m_{SR}) = r_T \tag{14}$$

As before it is possible to eliminate from the seven Equations 5, 6, 10–14 additional with $m_S^* = P_{ES}\, m_S$ the variables m_F, m_F^*, s_F, r_F, m_R, m_S, m_S^* and m_{SR}. The result is a quintic polynomial equation describing the translational active mRNA m_{TA} as a function of K_S, K_R, n, P_{EF}, P_{ES}, m_T, r_T and s_T, which can be numerically solved.

Table 1. Overview of the modules from the `ViennaRNA Package` used in the implementation of our translation initiation model. Manuals with more detailed descriptions can be found at www.tbi.univie.ac.at/~ronny/programs/<program_name>.html.

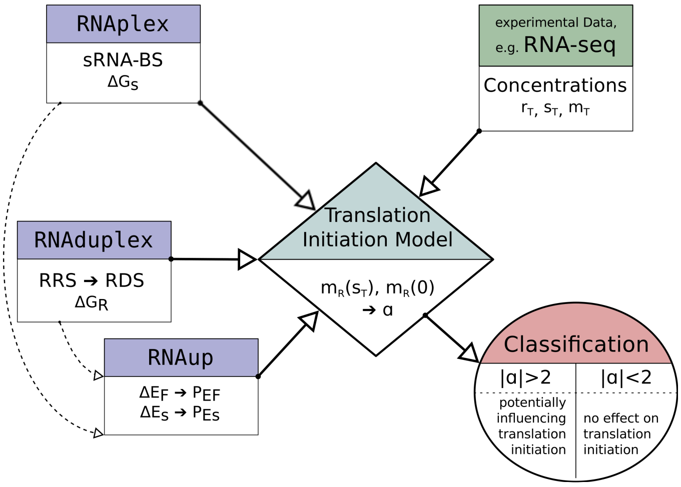
Program Name	Program Description	Reference
`RNAduplex`	Computes optimal structures upon hybridization of two RNA strands and the free energy of the resulting duplex. The calculation is simplified by allowing only inter-molecular base pairs.	[10]
`RNAplex`	Finds optimal sub-optimal target sites of a query RNA on an mRNA by computing secondary structures for their hybridization. Accessibility effects are included in an approximate manner, based on accessibility profiles computed by RNAplfold.	[13]
`RNAplfold`	Performs local folding of very long sequences, allowing only base pairs with a maximal span of L. It computes mean pair probabilities as well as accessibilities for every position i, averaging over all sequence windows of length W that contain i. The resulting accessibility profiles can be used, e.g., in RNAplex.	[14,15]
`RNAup`	Computes accessibilities, *i.e.*, the probability $P_u[i,j]$ that a sequence interval $[i,j]$ is unpaired, with an extension of the standard partition function approach for RNA secondary structure. This computation can also be conducted with constraints to force specified bases to remain unpaired, which allows us to compute accessibilities with- and without bound sRNA.	[11]

2.3. Model Implementation

The described model equations contain concentration data, s_T, m_T, r_T and n, which can be deduced from experiments (e.g., RNA-seq or tiling arrays) and equilibrium constants, K_S, K_R, P_{ES} and P_{EF}, which all can be calculated. To perform this calculations and solve the equations, we developed a software-wrapper that makes extensive use of programs included in the `ViennaRNA Package` [10].

A more detailed description of the programs used can be found in Table 1. Figure 2 illustrates the main work-flow of the model implementation.

Figure 2. Illustration of the work-flow for the classification of whether sRNA binding can influence the mRNA's translation initiation. `RNAplex` is used to calculate possible sRNA–mRNA interaction sites. `RNAduplex` calculates the ribosome–mRNA interaction, hence determining the position of the RRS and RDS, and the hybridization energy ΔG_R. The position of the RDS and the sRNA binding site (sRNA-BS) is used with `RNAup` to determine the exposing probabilities P_{EF} and P_{ES}. The concentrations of all reactants are deduced from RNA-seq data. All this information is integrated in the *Translation Initiation Model* to calculate the amount of mRNA that is bound by the initiation complex assuming the presence ($m_R(s_T)$) and the absence ($m_R(0)$) of sRNA. The ratio α of these serves as a descriptor to classify the potential of the sRNA to influence translation initiation.


RNAplex
sRNA-BS
ΔGS
experimental Data, e.g. RNA-seq
Concentrations
rT, sT, mT
Translation Initiation Model
mR(sT), mR(0) → α
RNAduplex
RRS → RDS
ΔGR
RNAup
ΔEF → PEF
ΔES → PES
Classification
|α|>2
|α|<2
potentially influencing translation initiation
no effect on translation initiation

The potential sRNA-BS are determined with `RNAplex`, considering the accessibility of potential binding sites on the sRNA and mRNA. The accessibility is calculated with `RNAplfold` (the `-W` and `-L` parameter are set to 200 and 150, respectively [24]). All sub-optimal binding sites up to a binding energy ΔG_S of $-7\ ^{kcal}/_{mol}$, which are at most 150 nt upstream to 20 nt downstream of the translation start site and at least 10 nt long (including inter-molecular bulges), are considered for follow-up evaluation of their potential to influence translation initiation. `RNAplex`-based target prediction results in sRNA-BS coordinates and the binding energy ΔG_S, which includes (in contrast to ΔG_R) the energy needed to make the binding sites accessible.

The search for the RRS is performed by `RNAduplex`, which calculates the energy and position of the optimal binding site between two given RNA molecules. `RNAduplex` only considers inter-molecular interactions. Intra-molecular base-pairs are ignored but inter-molecular bulges and internal loops are permitted [25]. The search space was set to -30 nt upstream to +3 nt downstream of the translation start site against the 10 nucleotides at the $3'$ end of the 16S rRNA. This provides the position of the RRS and the corresponding hybridization energy ΔG_R of the ribosome to the mRNA. From the RRS position the RDS position can be directly deduced to be $RRSstart$ to $RRSstart + 30\ nt$. The opening energy ΔE_F

of the RDS was calculated with `RNAup` [11], using a sequence stretch of ± 250 nucleotides around the RDS. From the opening energy ΔE_F the probability of being fully unfolded can be deduced from $P_{EF} = \exp(-\frac{\Delta E_F}{R \cdot T})$.

To calculate the sRNA influenced opening energy ΔE_S, `RNAup` is used again. However, this time a constraint folding approach is applied, which prevents bases interacting with the sRNA to participate in intra-molecular folding of the mRNA. Once again, the probability P_{ES}, that the complete RDS is unstructured, is given by $P_{ES} = \exp(-\frac{\Delta E_S}{R \cdot T})$.

The software wrapper is provided with the anti-RRS sequence, an sRNA sequence, an mRNA sequence with annotated translation start site and information about the concentrations of the reaction members. The reaction constants are determined as described above and fed into the corresponding equation system to solve the number of ribosome bound mRNA, hence translational active mRNA, in the presence of sRNA, $m_{TA}(s_T)$, and without sRNA, $m_{TA}(0)$. The corresponding equation is solved numerically applying Newton's method. In the case where RDS and sRNA-BS overlap, we can set $m_{TA} = m_R$. For each analyzed putative sRNA binding site, the signed ratio $\alpha = \frac{m_{TA}(s_T)}{m_{TA}(0)}$ (or $\alpha = -\frac{m_{TA}(0)}{m_{TA}(s_T)}$ if $m_{TA}(0) > m_{TA}(s_T)$) is returned as a measure of the sRNA induced change of translation initiation efficiency. We consider all mRNA whose translation initiation rate changes more than 2-fold ($|\alpha| > 2$) to be putatively regulated by the corresponding sRNA.

3. Simulation of Known sRNA–mRNA Interactions

To test our model, we simulated the effect of sRNA binding onto translation initiation for several well-described sRNA and their targets. Since the distinguishing characteristic of the presented approach is the possibility to qualify the regulatory effect of a proposed sRNA–mRNA interaction, the focus was set on all sRNA in *E. coli* for which experimentally validated cases of positive regulation are known, *i.e.*, DsrA, RprA, ArcZ, GlmZ and RyhB [26]. Thereby all confirmed interactions (positive as well as negative) of those sRNA were simulated (see Table 2).

The mRNA expression levels were estimated from publicly available deep sequencing data obtainable at the *Sequence Read Archive* (submission ID: SRA050648). Briefly, *E. coli MG1655* was grown in rich media, no rRNA depletion was performed prior to RNA-seq, 49,979,354 reads were produced [27]. The obtained reads were mapped onto *E. coli* genome (NC_000913), using segemehl [28] with default settings. The mapped reads were assigned to the corresponding protein coding and ncRNA genes, annotated in the Refseq database. If a read mapped n times equally well to the genome, we counted $^1/_n$ for each position. The counts for each gene were normalized for gene length and total read count (RPKM, Reads Per Kilobase of gene per Million mapped reads). The total number of 16S rRNA molecules within the cell is assumed to be 57,000 [20,23]. Thus, the RPKM values for each gene were further normalized by dividing by the sum of all seven 16S rRNA RPKM values and multiplied by 57,000. The resulting values are supposed to reflect the concentration ratios between the 16S rRNA and the mRNA molecules. 3833 genes were shown to be transcribed, 489 genes showed no transcription at all.

Table 2. The modeled changes in translation initiation rate for five sRNA. *Regulation Type* gives the experimentally shown behavior of the system. *Position (mRNA)* gives the calculated site of sRNA binding onto the mRNA relative to the start codon. *Hybridization Energy* gives the energy gained by the hybridization of the mRNA and the sRNA in $^{kcal}/_{mol}$. *Fold Change* α is the resulting value, according to the simulation, how much the initiation rate changes with and without sRNA.

sRNA	**mRNA**	**Regulation Type**	**Position (mRNA)**	**Hybridization Energy**	**Fold Change** α
dsrA	hns	repression	(−12)..+18	−22.9	−2.94
	rpoS	activation	(−126)..(−97)	−33	+2.99
rprA	rpoS	activation	(−133)..(−94)	−30.7	+2.11
arcZ	rpoS	activation	(−105)..(−81)	−23.3	+13.50
	sdaC	repression	(−13)..(−3)	−13	−2.90
	tpx	repression	—	—	—
glmZ	glmS	activation	(−40)..(−22)	−19.2	+26.23
ryhB	shiA	activation	(−59)..(−48)	−19.2	±1
	ufo/fur	repression	(−31)..(−18)	−13.1	−2.99
	cysE	repression	(−11)..+8	−27.0	−3.00
	frdA	repression	(−17)..+3	−24.5	−2.97
	iscS	repression	(−26)..+2	−23.7	−2.92
	dadA	repression	+9..+39	−29.2	−3.01
	sodB	repression	(−)21..+4	−18.8	−3.00
	sdhC	repression	(−28)..(−8)	−17.1	−2.87

At any given moment, about 80% of the ribosomes are actively engaged in translation [29], thus reducing the number of total ribosome r_T that are available for translation initiation to 11,400 per cell. Unfortunately, many of the sRNA are not expressed under the conditions of the RNA-seq experiment. We therefore used an ad-hoc estimate of sRNA concentrations (under conditions where the sRNA is active) and we set the ratio of sRNA and mRNA molecules to be $^2/_3$. This is motivated by the idea that, presuming a similar state of the transcriptome, inducing sRNA gene expression to a level of $^2/_3$ of the target gene should already yield a visible effect on the translation initiation rate of the target gene. To get a rough estimate of the scale of this ratio in bacteria, we examined all *E. coli* trans-acting ncRNA from the EcoCyc database [30] whether they were shown to be expressed in rich growth medium. Seven ncRNA genes fulfill this criterion, of which five are also described in terms of their targets (e.g., micM, mcaS, glmY, omrA and mgrR with a total of 12 targets). We calculated the $\frac{[sRNA]}{[mRNA]}$ ratios for all sRNA-target pairs from the normalized RPKM values and deduced the geometric mean of 0.84, close to our value 0.67 estimated from theoretical considerations.

To use the most realistic model of the mRNA possible, we reconstructed primary transcripts from a detailed analysis of the *E. coli* transcriptome [31], considering the experimentally validated operonic architecture and transcription start sites.

Based on this, thirteen out of fifteen experimental interactions could be modeled qualitatively correctly (Table 2). For tpx `RNAplex` does not find any potential ArcZ binding site ±200 nt around the ribosome docking site that has less than or equal to -7 $^{kcal}/_{mol}$ free energy. The calculated interaction between RyhB and shiA takes place from -59 nt to -48 nt. This is in contradiction to the binding site found experimentally, which is between position -76 nt and -27 nt upstream of the translation start site [32]. Applying this elongated binding site leads to a RDS accessibility change from $P_{EF} = 1.2 \times 10^{-5}$ to $P_{ES} = 7 \times 10^{-4}$.

4. Usage as Target Prediction Tool

The presented translation initiation model under the influence of sRNA binding has two new features. First, it integrates information about the transcript concentrations and the thermodynamic properties of the sRNA–mRNA and the ribosome–mRNA system. Second, it is possible to evaluate all putative binding sites for their capability to influence translation initiation. The first should be helpful in increasing the specificity, the latter should increase sensitivity, compared with existing target prediction methods.

To test the predictive power of our model, it was applied to predict all sRNA that can regulate RpoS translation, as well as all mRNA that are putatively regulated by RyhB.

4.1. Searching sRNA Controlling RpoS Translation

RpoS is an especially interesting gene because it was shown that it is activated by three different sRNA. RpoS is an alternative σ-factor that helps the RNA polymerase to recognize promoters of genes involved in stress response and secondary metabolism [33], thus making RpoS a central node in integrating information about the status of the cell. This is achieved by a variety of regulatory mechanisms on all levels. Beside, the known sRNA regulators of rpoS, it was suggested that other so far unknown sRNA may regulate rpoS translation [34,35].

All ncRNA from Refseq that are annotated neither as ribosomal nor tRNA (65 genes in total) were used to evaluate their potential effect on RpoS translation. An interaction is considered potentially functional if it causes more than ±2 fold change α, takes place at -150 nt to +20 nt from the translation start, and has an interaction length of at least 10 nt and a binding energy ΔG_S of at most -7 $^{kcal}/_{mol}$.

Six ncRNA fulfilled these criteria (Table 3). Beside the three above-mentioned known interactions, taken from [26], there are three additional sRNA genes with the potential to repress RpoS translation. All three of them have a higher, thus less favorable, hybridization energy, compared with the validated interactions. For OxyS it was already reported that oxyS over-expression decreases RpoS expression [36].

This analysis was also used to test how sensitive the results are to the chosen parameters. The same analysis was performed with different values for the ribosome occupancy ($n = 1 \text{ to } 100$) and for the concentration ratios $\frac{[sRNA]}{[mRNA]} = {}^{1}/_{2}, {}^{2}/_{3}, {}^{1}/_{1}, {}^{3}/_{2}, {}^{2}/_{1}, {}^{3}/_{1}, {}^{4}/_{1}$. For all of them the same potential regulators were predicted, except for $\frac{[sRNA]}{[mRNA]} = {}^{1}/_{2}$ where only dsrA and arcZ showed the potential to influence rpoS translation.

Table 3. The modeled changes in translation initiation rate. 65 ncRNA from *E.coli* were tested against rpoS mRNA. Six show a fold change greater than ±2. The table is sorted in ascending order according to their *Hybridization Energy*.

sRNA	**mRNA**	**Position (mRNA)**	**Hybridization Energy**	**Fold Change α**
dsrA	rpoS	$-126..-97$	-33.0	+3.0
rprA	rpoS	$-133..-94$	-30.7	+2.1
arcZ	rpoS	$-105..-81$	-23.3	+13.5
omrA	rpoS	$-27..-9$	-21.3	-2.8
ryjA	rpoS	$-22..-8$	-17.4	-2.8
oxyS	rpoS	+17..+27	-13.1	-2.8

4.2. Searching mRNA Controlled by RyhB

RyhB is a 90 nt long sRNA that plays an important role in cell homeostasis. Under conditions of iron starvation, RyhB is expressed and reduces the translation of non-essential iron-using proteins [37].

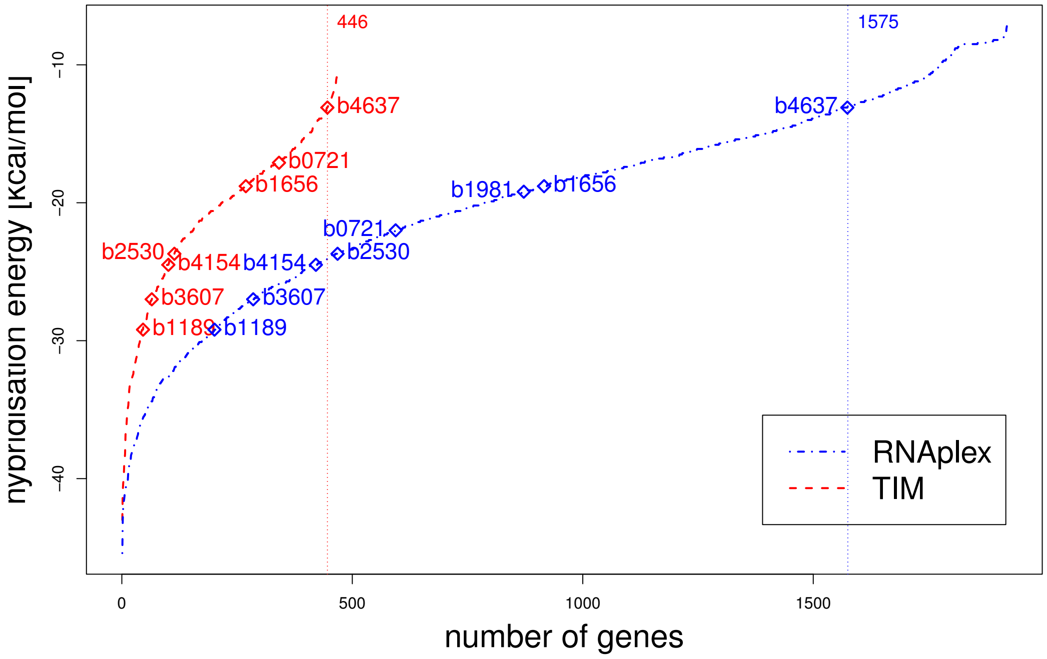
Potential binding sites of RyhB on all 4146 protein coding genes annotated in NCBI Refseq were calculated with `RNAplex`. 1921 genes, including all eight known targets, have an RyhB binding site with a binding energy $\Delta G_S \leq -7$ $^{\mathrm{kcal}}/_{\mathrm{mol}}$ in the vicinity of their translation start (-150 nt to +20 nt).

Sorting the most favorable binding sites in this neighborhood according to their hybridization energy, without post evaluation with our translation initiation model, results in the eight interactions described in literature among the 1575 most stable interactions (Figure 3).

After applying our translation initiation model and removing all interactions that seem to lack the potential to change the translation initiation rate more than ±2 fold, only 446 binding sites had a more stable hybridization energy than the least stable known interaction (b4637 with -13.1 $^{kcal}/_{mol}$). In total 467 genes seemed to be potentially targeted by RyhB (Figure 3). Here shiA (b1981), a well documented activation target of RyhB, is no longer detected (see Section 3). A more detailed inspection of one particular putative RyhB target is given in the supplementary material (Section 1).

We compared the found 467 genes with experimental results from micro-array analysis with an inducible ryhB gene [38]. A general drawback of this kind of experiment is the difficulty to distinguish between directly and indirectly regulated genes. The authors tried to circumvent this by reducing the time span between RyhB induction and the assay to 15 min. This time could be still too long considering their own results for the gene exbBD, which, although most probably an indirect regulated gene, showed already after 7.5 min a significant drop in mRNA abundance. To identify genes regulated by fur, which itself is regulated by RyhB, the assay was compared with a fur^- mutant. In spite of this precautionary measure, the possibility that another transcription factor is RyhB controlled cannot be ruled out, hence the targets found can still be indirectly regulated by RyhB. In [38], 56 gene targets from 18 different operons could be identified as being regulated by RyhB, whereas in our analysis, in 12 out of 18 operons (∼67%) we find at least one gene that is regulated in the same sense than observed in the micro-array experiment.

Figure 3. The distribution of hybridization energy. The blue curve shows the minimal hybridization energy for each gene with a calculated binding site from -150 nt upstream to +20 nt downstream of the translation start site and $\Delta G \leq -7\ ^{\mathrm{kcal}}/_{\mathrm{mol}}$. The experimental validated genes are marked with $\diamond$. In contrast, the red curve shows the hybridization energy for all genes that are potentially altered in their expression by RyhB, according to our Translation Initiation Model (TIM).



It is worth noting that it is not always the energetically most favorable binding site within our search region of -150 nt to +20 nt around the translation start site, which has the strongest effect on translation initiation in our model. For example, RyhB can bind sdhC (b0721) -149 nt in front of the translation start with an hybridization energy of $-22.0\ ^{kcal}/_{mol}$. According to our model, this has a negligible effect on translation initiation of $\alpha = +1.0007$. The energetically less favorable binding site with $-17.1\ ^{kcal}/_{mol}$, which overlaps the ribosome docking site, has a significant effect of $\alpha = -2.87$.

Although the pool of putative targets could be decreased by our binding site evaluation, 467 targets ($\sim$10 % of all genes) still seem implausible. At the moment, a comprehensive set of confirmed direct RyhB targets is still lacking, which would enable a detailed analysis of the specificity and sensitivity of our modeling approach. To get at least an idea of the significance of our results, we tested those 467 genes for the enrichment of certain functions described with Gene Ontology terms [39] using a web-based tool (Database for Annotation, Visualization and Integrated Discovery (`DAVID`) [40]). This revealed that $\sim$5 % of the putative targets are associated with the GO term *anaerobic respiration* and $\sim$10 % with the term *iron ion binding* (see Table 4). The p-values of this enrichment are 1.0×10^{-12} and 7.3×10^{-10}, respectively. This is in perfect agreement with the role of RyhB in the cell, indicating that the regulon of RyhB is indeed much larger than the experimentally validated eight targets.

Table 4. Gene Ontology term enrichment analysis of 467 genes that appeared to be potentially influenced by RyhB. The analysis was performed with `DAVID`. The gene list is highly enriched with genes associated with the GO terms *anaerobic respiration* and *iron ion binding*. The *p-value* expresses the likelihood of the observed enrichment happening by chance. *Count* and % give the number of genes and the percentage of the whole list of 467 genes associated with the corresponding GO term.

GO name space	GO Term	Count	%	*p*-value
biological process	GO:0006091 generation of precursor metabolites and energy	49	10.5 %	1.0×10^{-12}
biological process	GO:0009061 anaerobic respiration	22	4.7 %	9.5×10^{-12}
molecular function	GO:0043169 cation binding	96	20.6 %	2.5×10^{-10}
molecular function	GO:0046872 metal ion binding	94	20.2 %	2.8×10^{-10}
molecular function	GO:0043167 ion binding	96	20.6 %	3.4×10^{-10}
molecular function	GO:0005506 iron ion binding	45	9.7 %	7.3×10^{-10}

5. Discussion

We presented a method to evaluate the capability of predicted sRNA–mRNA interactions in interfering translation initiation. We successfully simulated the effect of five *Escherichia coli* sRNA onto their experimentally validated targets. Furthermore, we used our method to predict potential regulators of RpoS and potential targets of RyhB. The latter was compared with target prediction without post-processing. Applying our translation initiation model reduces the list of successfully predicted known targets from eight to seven. At the same time, the number of potential targets is reduced from 1921 genes to 467 genes.

A further novelty of our approach is the possibility to distinguish between translation activation and repression for the predicted sRNA–mRNA interaction. While we show the usefulness of calculated fold changes in the formation of initiation complexes (α values), there remain reasons to be cautious with a quantitative interpretation of α values. For example, our model considers only one binding site at a time, and therefore does not model the competition of several mRNA for an sRNA. Moreover, the actual kinetics might be more important than the equilibrium state, especially because of the fact that bacterial translation initiation already occurs co-transcriptional, changing the chronology of binding sites becoming available, which can drastically change the kinetic behavior of the system from the equilibrium state. Finally, translation initiation is a highly stochastic process occurring in bursts [23], which is not considered in the presented model. Considering this, we do not think that the α values can serve as suitable classifier to rank the reliability of predicted targets. Nevertheless, we could show that a mere binding energy based ranking leads to a significant enrichment of known targets within the top ranked genes, after evaluation of the binding sites. For the application of our model to the sRNA RyhB, there are no known targets within the top 125 ranked genes, according to a mere interaction based prediction. In contrast, after evaluating the putative interactions with our translation initiation model, we find 4 known targets within the top ranked 125 genes.

Our target prediction approach is the first to explicitly model the concentration dependence of sRNA–mRNA binding. With the advances in high throughput transcriptome quantification, such as RNA-seq or genomic tiling arrays [41], more data on mRNA expression levels are becoming available. Unfortunately, these data often do not include sRNA or are not measured under conditions relevant for sRNA regulation. We tried to find a compromise for this by deducing the concentration ratios between mRNA and ribosome from biological experiments, but assumed the ratio $\frac{[sRNA]}{[mRNA]}$ to be $2/3$. In the near future, when more expression data for different species and different conditions will be publicly available or cheaper to produce, this problem might be overcome.

The role of the RNA chaperon Hfq is not considered in our model. Hfq is thought to enable sRNA-based translation regulation either by (1) protecting the sRNA from RNase E degradation, (2) recruiting RNase E to degrade the Hfq·mRNA·sRNA complex, or (3) facilitating the interaction between sRNA and mRNA [42]. The first would change sRNA abundance, which we avoid by assuming an effective sRNA concentration in the first place. The second mechanism, where Hfq mediates mRNA degradation, is ignored in our model which exclusively describes the sRNA effect on translation initiation. For the last mentioned mechanism, Hfq works as a chaperon, changing the kinetics of sRNA–mRNA interaction. It was shown that sRNA·mRNA complexes established this way remain stable after Hfq removal [43]. This implies that regarding the thermodynamic equilibrium state may be sufficient to detect Hfq dependent targets. An extension of our model, including effects of Hfq, is possible, but would require more knowledge about the strength and specificity of RNA–Hfq interactions.

The discrepancy in the number of confirmed interactions from biological experiments and from computational screens is puzzling. To our knowledge, the most comprehensive investigation of an sRNA regulon was published by Sharma *et al.* [44]. There, a genome-wide experimental approach and bioinformatic target prediction was combined. The regulon of GcvB in *Salmonella thyphimurium* could be enlarged to 54 genes, which corresponds to 45 different cistrons, of which 21 could be individually confirmed. We agree with the authors that this is most likely not the end of the line. Due to the fact that so far most genomic screens are solely based on changes in mRNA concentrations, which do not have to go along with translational regulation, some targets could be still missed. Furthermore, technical difficulties (e.g., read out methods) can increase the false negative rate.

Conferring this analysis to the situation of RyhB in *Escherichia coli*, together with the fact that our prediction method found 45 new targets associated with the molecular function “iron ion binding”, suggests that the regulon of RyhB is indeed much larger. Besides, it shows that our bioinformatic approach of blending RNA interaction with translation initiation is a promising tool for sRNA target prediction.

We plan to provide the described approach to the scientific community as a web-based service incorporated into the RNApredator [45] target prediction web-server (http://rna.tbi.univie.ac.at/RNApredator/) as a post-processing analysis.

6. Conclusions

From our point of view, computational and experimental techniques each have their advantages and disadvantages. For a complete understanding of the role of sRNA in the bacterial cell, computational and experimental biologists should rethink and enlarge their repertoire of techniques. We hope that the presented approach serves to this end.

Acknowledgments

We thank Sven Findeiss and Peter Kerpedjiev for commenting and proofreading this manuscript.

This work was partly funded by the Austrian GEN-AU project "Bioinformatics-Integration-Network III" and the Austrian Science Fund (FWF) project "In silico annotation of noncoding RNAs and their targets" (AF 0430511).

Conflict of Interest

The authors declare no conflict of interest.

References

1. Mizuno, T.; Chou, M.; Inouye, M. A unique mechanism regulating gene expression: Translational inhibition by a complementary RNA transcript (micRNA). *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 1966–1970.
2. Raghavan, R.; Groisman, E.; Ochman, H. Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res.* **2011**, *21*, 1487–1497.
3. Backofen, R.; Hess, W. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol.* **2010**, *7*, 33–42.
4. Waters, L.; Storz, G. Regulatory RNAs in bacteria. *Cell* **2009**, *136*, 615–628.
5. Witkos, T.; Koscianska, E.; Krzyzosiak, W. Practical aspects of microRNA target prediction. *Curr. Mol. Med.* **2011**, *11*, 93.
6. Urban, J.; Vogel, J. Translational control and target recognition by *Escherichia coli* small RNAs *in vivo*. *Nucl. Acids Res.* **2007**, *35*, 1018–1037.
7. Sharma, C.; Vogel, J. Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr. Opin. Microbiol.* **2009**, *12*, 536–546.
8. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
9. Tjaden, B. TargetRNA: A tool for predicting targets of small RNA action in bacteria. *Nucl. Acids Res.* **2008**, *36*, W109–W113.
10. Lorenz, R.; Bernhart, S.; zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.; Hofacker, I. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26.
11. Mückstein, U.; Tafer, H.; Hackermüller, J.; Bernhart, S.; Stadler, P.; Hofacker, I. Thermodynamics of RNA–RNA binding. *Bioinformatics* **2006**, *22*, 1177–1182.
12. Busch, A.; Richter, A.; Backofen, R. IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **2008**, *24*, 2849–2856.
13. Tafer, H.; Amman, F.; Eggenhofer, F.; Stadler, P.; Hofacker, I. Fast accessibility-based prediction of RNA–RNA interactions. *Bioinformatics* **2011**, *27*, 1934–1940.

14. Bernhart, S.; Hofacker, I.; Stadler, P. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **2006**, *22*, 614–615.
15. Bompfünewerer, A.F.; Backofen, R.; Bernhart, S.H.; Hertel, J.; Hofacker, I.L.; Stadler, P.F.; Will, S. Variations on RNA folding and alignment: Lessons from Benasque. *J. Math. Biol.* **2008**, *56*, 129–144.
16. Fröhlich, K.; Vogel, J. Activation of gene expression by small RNA. *Curr. Opin. Microbiol.* **2009**, *12*, 674–682.
17. Holmqvist, E.; Reimegård, J.; Sterk, M.; Grantcharova, N.; Römling, U.; Wagner, E. Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J.* **2010**, *29*, 1840–1850.
18. Valentin-Hansen, P.; Eriksen, M.; Udesen, C. MicroReview: The bacterial Sm-like protein Hfq: A key player in RNA transactions. *Mol. Microbiol.* **2004**, *51*, 1525–1533.
19. Laursen, B.; Sørensen, H.; Mortensen, K.; Sperling-Petersen, H. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 101–123.
20. Na, D.; Lee, S.; Lee, D. Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst. Biol.* **2010**, *4*, 71.
21. de Smit, M.; Van Duin, J. Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 7668–7672.
22. Salis, H.; Mirsky, E.; Voigt, C. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **2009**, *27*, 946–950.
23. Xie, X.; Choi, P.; Li, G.; Lee, N.; Lia, G. Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* **2008**, *37*, 417–444.
24. Lange, S.; Maticzka, D.; Möhl, M.; Gagnon, J.; Brown, C.; Backofen, R. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucl. Acids Res.* **2012**, *40*, 5215–5226.
25. Schurr, T.; Nadir, E.; Margalit, H. Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucl. Acids Res.* **1993**, *21*, 4019–4023.
26. Storz, G.; Vogel, J.; Wassarman, K. Regulation by small RNAs in bacteria: Expanding frontiers. *Mol. Cell* **2011**, *43*, 880–891.
27. Giannoukos, G.; Ciulla, D.; Huang, K.; Haas, B.; Izard, J.; Levin, J.; Livny, J.; Earl, A.; Gevers, D.; Ward, D.; *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **2012**, *13*, r23.
28. Hoffmann, S.; Otto, C.; Kurtz, S.; Sharma, C.; Khaitovich, P.; Vogel, J.; Stadler, P.; Hackermüller, J. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **2009**, *5*, e1000502.
29. Bremer, H.; Dennis, P. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli Salmonella: Cell. Mol. Biol.* **1996**, *2*, 1553–1569.
30. Keseler, I.; Collado-Vides, J.; Santos-Zavaleta, A.; Peralta-Gil, M.; Gama-Castro, S.; Muñiz-Rascado, L.; Bonavides-Martinez, C.; Paley, S.; Krummenacker, M.; Altman, T.; *et al.* EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucl. Acids Res.* **2011**, *39*, D583–D590.

31. Cho, B.; Zengler, K.; Qiu, Y.; Park, Y.; Knight, E.; Barrett, C.; Gao, Y.; Palsson, B. The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotechnol.* **2009**, *27*, 1043–1049.
32. Prévost, K.; Salvail, H.; Desnoyers, G.; Jacques, J.; Phaneuf, É.; Massé, E. The small RNA RyhB activates the translation of shiA mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. *Mol. Microbiol.* **2007**, *64*, 1260–1273.
33. Maciąg, A.; Peano, C.; Pietrelli, A.; Egli, T.; de Bellis, G.; Landini, P. *In vitro* transcription profiling of the σ^S subunit of bacterial RNA polymerase: Re-definition of the σ^S regulon and identification of σ^S-specific promoter sequence elements. *Nucl. Acids Res.* **2011**, *39*, 5338–5355.
34. Ruiz, N.; Silhavy, T. Constitutive activation of the *Escherichia coli* Pho regulon upregulates rpoS translation in an Hfq-dependent fashion. *J. Bacteriol.* **2003**, *185*, 5984–5992.
35. Peterson, C.; Carabetta, V.; Chowdhury, T.; Silhavy, T. LrhA regulates rpoS translation in response to the Rcs phosphorelay system in *Escherichia coli*. *J. Bacteriol.* **2006**, *188*, 3175–3181.
36. Zhang, A.; Altuvia, S.; Tiwari, A.; Argaman, L.; Hengge-Aronis, R.; Storz, G. The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein. *EMBO J.* **1998**, *17*, 6061–6068.
37. Salvail, H.; Massé, E. Regulating iron storage and metabolism with RNA: An overview of posttranscriptional controls of intracellular iron homeostasis. *Wiley Interdiscip. Rev.: RNA* **2011**, *3*, 26–36.
38. Massé, E.; Vanderpool, C.; Gottesman, S. Effect of RyhB small RNA on global iron use in *Escherichia coli*. *J. Bacteriol.* **2005**, *187*, 6962–6971.
39. Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H.; Cherry, J.; Davis, A.; Dolinski, K.; Dwight, S.; Eppig, J.T.; *et al.* Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29.
40. Huang, D.B.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2008**, *4*, 44–57.
41. Mäder, U.; Nicolas, P.; Richard, H.; Bessières, P.; Aymerich, S. Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Curr. Opin. Biotechnol.* **2011**, *22*, 32–41.
42. Vogel, J.; Luisi, B. Hfq and its constellation of RNA. *Nat. Rev. Microbiol.* **2011**, *9*, 578–589.
43. Moll, I.; Leitsch, D.; Steinhauser, T.; Bläsi, U. RNA chaperone activity of the Sm-like Hfq protein. *EMBO Rep.* **2003**, *4*, 284–289.
44. Sharma, C.; Papenfort, K.; Pernitzsch, S.; Mollenkopf, H.; Hinton, J.; Vogel, J. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol. Microbiol.* **2011**, *81*, 1144–1165.
45. Eggenhofer, F.; Tafer, H.; Stadler, P.; Hofacker, I. RNApredator: Fast accessibility-based prediction of sRNA targets. *Nucl. Acids Res.* **2011**, *39*, W149–W154.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

Part IV

Discussion

10 Discussion

Once the relevance of sRNA based gene regulation in bacteria was endorsed, the scientific community worked specifically towards improving the techniques to annotate sRNA genes. Only in a second phase, in the last few years the efforts to characterization the found genes were stepped up, e.g. in [150, 209, 218]. Nevertheless, detailed functional descriptions of sRNA genes still are more of anecdotal nature, if compared to the number of annotated sRNA genes. In the coming years, the characterization of the underlying regulatory networks should shift into the focus of research. To describe this arising network the single participants, namely the sRNA and the transcripts, have to be identified first. In the sense of a network view these constitute the vertices in the network graph. On the sRNA side, the genes and their precise start and end points can be determine with a combination of in silico and in vitro analysis. The detection of protein coding regions is a well established technique. In contrast, the toolbox to determine the exact transcript boundaries and architecture is much less well equipped.

The edges of the regulatory network graph represent different interaction types between the members. Interactions between protein coding genes, in which one gene product influences the activity of another gene's transcription, are subject of intensive research since many decades. For the bacterial model organism *E. coli* the database `RegulonDB` [67] aims to collect all this interactions, which are mostly transcription factor to target gene interactions. For sRNA based interactions the repertoire of applicable techniques is much smaller.

To complement a gene regulatory network with sRNA mediated regulations for one particular species, it would be necessary to detect all sRNA genes and identify all the influenced targets of the sRNA. This is a Herculean task, which is not even remotely achieved for any species. Nevertheless, this should be the ultimate goal of the bacterial ncRNA research community. Therefore, new ideas and new tools for the characterization of single sRNA–mRNA interactions and whole regulatory networks are needed.

Within this thesis, several new tools were presented which aim to close the gaps in a purposeful analysis pipeline of sRNA mediated gene regulation. In the following, a work flow using these tools is proposed. Followed by a display of how these tools could be improved to further enhance the performance of the analysis. Finally, the potential of a straight forward and easy to use post-transcriptional gene regulation network analysis in the applied field of medicine and biotechnology is discussed.

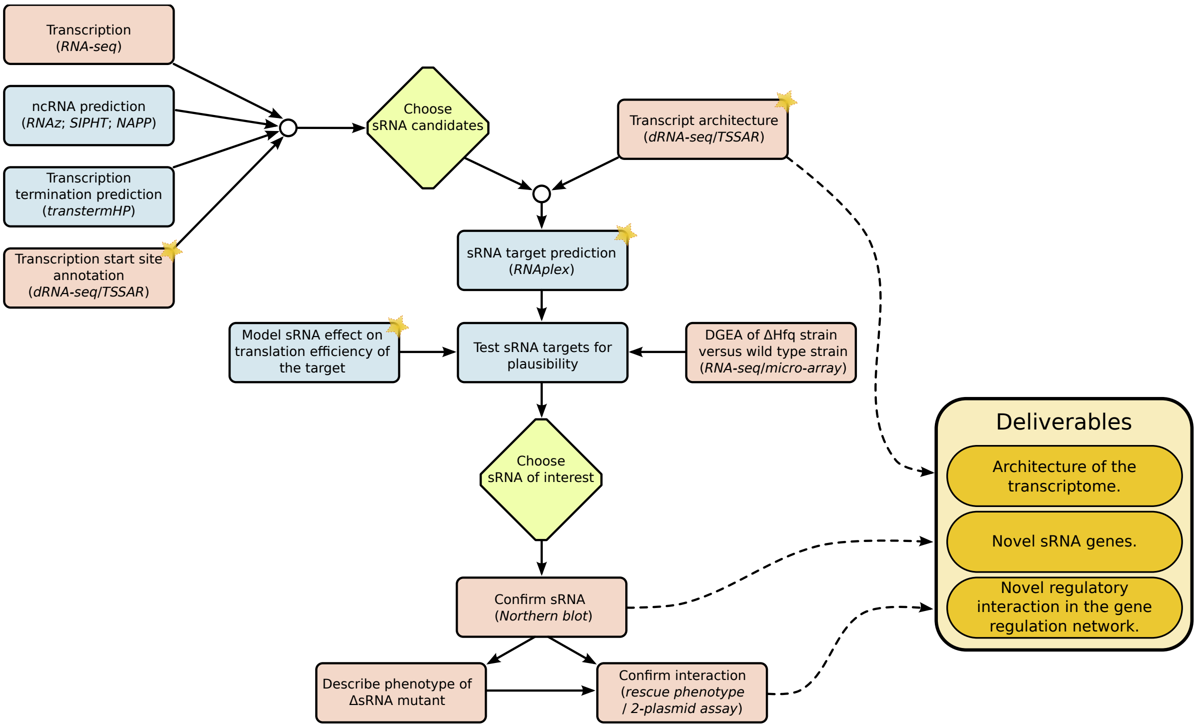

Transcription
(*RNA-seq*)
ncRNA prediction
(*RNAz; SIPHT; NAPP*)
Transcription termination prediction
(*transtermHP*)
Transcription start site annotation
(*dRNA-seq/TSSAR*)
Choose sRNA candidates
Transcript architecture
(*dRNA-seq/TSSAR*)
sRNA target prediction
(*RNAplex*)
Model sRNA effect on translation efficiency of the target
Test sRNA targets for plausibility
DGEA of ΔHfq strain versus wild type strain
(*RNA-seq/micro-array*)
Choose sRNA of interest
Confirm sRNA
(*Northern blot*)
Describe phenotype of ΔsRNA mutant
Confirm interaction
(*rescue phenotype / 2-plasmid assay*)
Deliverables
Architecture of the transcriptome.
Novel sRNA genes.
Novel regulatory interaction in the gene regulation network.

Figure 10.1: A propose work-flow to reveal the sRNA mediated gene regulatory network for a species of interest. Red boxes represents in vivo and in vitro analysis steps. Blue boxes describe in silico analysis steps. Star marked boxes involve tools developed in the course of this thesis.

10.1 Work-flow

A complete work-flow tackling the challenge of describing the sRNA based gene regulation can utilize all three presented tools to close critical gaps and reduce the downstream workload significantly. Exemplarily, the proposed work-flow is illustrated using *Bordetella pertussis* as a showcase[1].

To get a general idea of the importance of sRNA regulation in *B. pertussis* a ΔHfq mutant is constructed and examined by RNA-seq and micro-array for expression differences in the transcriptome. Since many sRNA are only functional with the assistance of Hfq, the knock out of this important RNA chaperon can be seen as a general cancellation of functional sRNA–mRNA interactions. By this, it can be shown that many genes, especially genes involved in

---

[1] A very similar analysis, which aims to reveal the role of sRNA based gene regulation in the establishment of virulence of the pathogen *Bordetella pertussis* is applied in an ongoing project, which is done in close cooperation with Dr. Branislav Večerek from the Institute of Microbiology, Academy of Sciences of the Czech Republic, v.v.i., and Dr. David Hot from the Center of Infection and Immunity of Lille, Institut Pasteur de Lille. I would like to express my gratitude to both of them for this fruitful and pleasurable collaboration. Chapters 7 and 9 present tools which were directly developed to cope with challenges emerging from this project.

virulence processes, are differentially expressed, which serves as an encouragement for a more detailed inspection of sRNA regulation. Therefore, the genome is systematically screened for sRNA genes. This is approached with different techniques. For one, computational approaches, such as RNAz, SIPHT and NAPP, are applied. On the other side, RNA-seq data are used to detect transcripts which have not been annotated so far and lack a long ORF, ruling out the possibility of being a newly found protein coding gene. The results from this diverse approaches show only a moderate overlap. To come up with a small and reliable set of potential sRNA genes all the information, combined with in silico predicted transcription termination hairpin signals and TSS deduced from dRNA-seq analysis, has to be manually screened. Each putative sRNA locus is assessed and finally only the most credible ones are considered for further analysis.

To characterize the newly found putative sRNA genes, the targets of this regulators are of pivotal interest. Therefore, first the potential binding partners, namely their target transcripts, have to be characterized. This was approached by a dRNA-seq experiment which was interpreted with TSSAR, see Chapter 7. The information on experimentally deduced TSS is also used to revisit the predicted ncRNA and reduce the list to only signals which seem to be independently transcribed from a own promoter, at least in the growth condition examined.

With the precise transcript structure as the basis, `RNAplex` is used to calculate potential binding interactions between the sRNA and the mRNA, see Chapter 8. First, only the computed energies are used to assign trustworthiness to the interactions. In a further step, each potential interaction is used to model the net effect on translation initiation of the sRNA binding, see Chapter 9. This information can be combined with the data from the ΔHfq transcriptome to examine whether the calculated effect is in agreement with the effect deduced from the differential gene expression analysis. This further dramatically shrinks the plausible putative sRNA mRNA interactions.

From there on, only sRNA potentially regulating genes involved in the host pathogen interaction are further considered. This small set of putative sRNA genes are thoroughly tested by northern blot analysis with different probes to confirm their expression and to determine the boundaries of the transcripts. Subsequently, mutants are constructed by altering the mRNA binding site on the sRNA. This mutants can be used to evaluate the phenotype concerning their virulence and their transcriptome. In the next step the phenotypes are attempted to be rescued by introducing complementary mutations in the sRNA binding site on the mRNA. This way, the interaction can be shown to be functional and the mechanism works indeed dependent on physical base pairing between the two RNA, which is a hallmark of sRNA based gene regulation.

This thorough examination results (if successful) in three major outcomes: (i) A detailed description of the architecture of the transcriptome with transcription start sites and operonic structure. (ii) New sRNA gene annotation (corresponds to new vertices in the gene regulatory network). (iii) New regulatory interactions between sRNA and mRNA genes (corresponds to new edges in the gene regulatory network). This leads to a better understanding of how bacteria cope with different environmental conditions and might bear new possibilities to intervene, for biotechnical or medical purpose, with the default genetic program (Fig. 10.1).

10.2 Improvements & Outlook

In the course of this thesis new tools, helpful to characterize sRNA based regulatory network in bacteria, were presented. Automatic and statistical sound TSS annotation (i.e. TSSAR), fast and accurate RNA interaction prediction (i.e. `RNAplex`), and modeling the effect of sRNA binding on mRNA translation closed gaps in the analysis pipeline to efficiently characterize the role of sRNA in bacteria cells. Nevertheless, none of the presented tools already exhausts all the given potential. All of them can still be either refined or applied to more general cases.

TSSAR proved to be an accurate, automated analysis of dRNA data. Especially the provided web service, with several additional analysis features, is optimized for its usability and is already[2] used by the transcriptomics community. Its novel scheme to preprocess the raw data, which can be quite large, on the client's computer and transfer only the essential and compressed information to the provided server resources, can serve as a role model. This scheme enables to outsource more resource intensive analysis to central hubs with the appropriate infrastructure and know-how, minimizing the training requirements for the end user and avoiding huge data traffic at the same time. This architecture might be applied for similar task more often in the near future, opening centralized web services to a broader application spectrum, especially in the field of data intensive, high-throughput analysis.

Nevertheless, TSSAR has still potential to be significantly improved. First of all, theoretically the presented method is kept universal enough to be applied to different RNA-seq protocols, which aim to enrich certain positions in a library compared to a background library. For example, beside the TEX using dRNA-seq protocol, there are alternative approaches proposed, e.g. using the enzyme TAP [251], for which TSSAR could be employed for the automated data analysis. Although they differ in the method details, eventually two libraries are produced by both methods, whose position wise significant enrichment has to be calculated.

In the same spirit, TSSAR could be used to tap another sort of information hidden in dRNA-seq data. Potentially, the protocol can be used not only to rich primary RNA 5' ends, but also RNA processing sites. If an RNA is preferentially cut at a certain position by an RNase, within the untreated library this 5' end can be expected to be overrepresented compared to randomly

---

[2] Prior printed publication (as at October, 2013).

introduced RNA ends formed by the mechanic fragmentation of the RNA in the course of RNA-seq library preparation [250]. But since it is not protected by an 5' triphosphate, it should be depleted in the treated library compared to the untreated library. In principle TSSAR could be applied to detect such secondary 5' end corresponding to RNA processing sites. To date this assumption lacks experimentally confirmation. If it holds, this would create an elegant and useful technique to detect RNA processing sites.

A further so far not tested putative field of application for TSSAR is a differential promoter usage analysis (DPUA). The activity of different promoters for the same gene under different growth conditions could be examined by a two step process. First, the TSS are annotated within each sample condition, by comparing the TEX treated with the untreated library. In a second step the same algorithm is used to compare the TEX treated library of one sample with the TEX treated library of the other sample. The union of step one, should provide a more detailed map of transcription start sites. The overlap of step one and two should provide the sites which are differentially used due to the different growth conditions. So far this application emanate from theoretical consideration, whose usefulness in practice remains to be proven.

A field for improvement of TSSAR will be the incorporation of multi-sample dRNAseq experiments to further increase sensitivity and specificity of TSSAR. So far, sample replicas are only pooled to increase sequencing depth and analyzed as one sample. All signals below a specified thresholds are ignored. In an improved TSSAR version each sample would be preferable analyzed on its own and only eventually the information is merged into a cumulative proposition. In such a way, also weak, not significant signals can be declared authentic if they show to be reproducible across multiple replicas. So far the majority of dRNAseq experiments are conducted with only very few replicas, mostly due to financial constraints. Nevertheless, this might change in the near future, making a multi-sample aware version advantageous.

Similar to the last consideration, the incorporation of multi-species comparison [90] could on one hand further increase the prediction sensitivity, and on the other hand, be a valuable tool to compare related species, strain or growth conditions for differences in their transcriptional regulation. This might provide valuable insights into strain-specific promoter usage. Hence, it could explain distinct phenotypes for very closely related genotypes [58].

RNAplex is a distinguished advancement in RNA-RNA interaction prediction. It combines the accurateness of resource intensive approaches, such as `RNAup`, with the speed of genome screening programs, such as `RNAhybrid`.

One critical shortcoming of `RNAplex` in its current form is the inability to find so called multiple kissing hairpins. This mode of interaction is frequently found between sRNA and mRNA. Thereby, the two molecules bind each other in two, or theoretically more, distinct regions, separated by a region with intra-molecular structures [7]. Often, each of the binding region alone are marginally stable. Only when both are evaluated altogether the structure is recognized as a prominent inter-molecular binding. The exhaustive calculation of all possible conformations of this type is computational very expensive [2, 97, 191]. Fortunately, `RNAplex` offers a straight forward heuristic approach. The ability of `RNAplex` to compute suboptimal interactions can be exploited therefore. Having a list of potential binding regions ready at hand, allows to screen putative combinations whether they are compatible with each other, or if they exclude each other sterically. From there, the en bloc result can be calculated from the sub-results. The only remaining challenge to avoid gaming away the out standing speed of `RNAplex`, is the definition of smart rules, which can efficiently distinguish between exclusive and combinable sub-binding sites with as little computational effort as possible. At the same time the rules must be strict enough to avoid biological implausible combination.

Translation initiation model with the influence of sRNA is one of the first attempts to simulate the effect of a regulator onto gene expression with the explicit purpose to predict the behavior of the system. It is based upon a rather simple model how sRNA, mRNA and ribosome interact. Hence, the model can be adapted at several points to cover additional details of the underlying physical processes.

First and foremost, it would be desirable to include the contribution of Hfq into the model. At the moment, in silico modeling of Hfq RNA interaction is not feasible, although there are first results providing kinetic parameter for this reaction, e.g. [56]. However, it is already possible to estimate a Hfq contribution in resolving adverse structures or facilitating sRNA binding, by considering potential Hfq binding sites. In this sense, an sRNA mRNA interaction can be supplied with an additional energy term if the binding motif A-R-N is present in several copies at the mRNA close to the predicted sRNA binding site [14]. Furthermore, the accessibility of the binding site itself can be incorporated into this energy bonus. On one hand, such a strategy would lead away from the more physical based simulation applied at the moment. On the other hand, the prediction accuracy might be significantly improved.

10.3 Future Application

As for any fundamental research advancement the question of its application imposes itself. There are two important fields which can profit from a deeper understanding of the underlying gene regulatory network.

10.3.1 Medicine

On the one hand, knowing how pathogens adapt their gene expression in the course of an infection might reveal potential targets to intervene with this process, for a better outcome in the treatment of an apparent infection. For several human pathogens it was shown that sRNA are directly involved in the establishment of infection [34, 253] and the modulation of the host immun response [36]. Recently, in multi-drug resistant *Staphylococcus aureus*, small RNA were shown to be specifically induced after antimicrobial exposure. It seems that, in contrast to protein coding genes, whose expression profiles cluster with strain or growth phase conditions, sRNA genes seem to be predominantly linked to antibiotic exposure, including sRNA responses which are specific for particular antibiotics [95].

10.3.2 Biotechnology

On the other hand, in biotechnological production there is strong economic concern to optimize the whole metabolic network for maximal outcome. Thus, the regulation network must be understood and modified accordingly. Therefore, concentrated efforts are made to expose the sRNA based regulation network in biotechnological important species, such as cyanobacteria [238], methane-producing archaea [33], and clostridium [38, 39]. Understanding of sRNA based regulation does not only help in manipulating naturally evolved systems but can also be used to design de novo regulatory circuits [189]. This opens exciting possibilities to construct synthetic biologic systems with exactly the properties needed.

Part V

Appendices

A List of Abbreviations

Table A.1: List of Abbreviations

aDB	Anti-downstream box
CDS	Coding DNA sequence
ChIP	Chromatin immunoprecipitation
DB	Downstream box
DGEA	Differential gene expression analysis
DNA	Desoxyribonucleic acid
DPUA	Differential promoter usage analysis
dRNA-seq	Differential RNA sequencing
EF	Elongation factor
GFP	Green fluorescence protein
GOI	Gene of interest
IF	Initiation factor
MEA	Maximal expected accuracy (structure)
MFE	Minimum free energy (structure)
mRNA	Messenger RNA
ncRNA	Non-coding RNA
OGT	Optimal growth temperature
ORF	Open reading frame
qPCR	Quantitative polymerase chain reaction
RACE	Rapid amplification of cDNA ends
RBS	Ribosome binding site
RF	Release factor
RNA	Ribonucleic acid
RNAP	RNA polymerase
rRNA	Ribosomal RNA
S	Svedberg (unit for sedimentation rate)
SD	Shine-Dalgarno sequence
sRNA	Small RNA
TIR	Translation initiation region
tRNA	Transfer RNA
TSS	Transcription start site
UTR	Untranslated region

B List of Tables

3.1 Rfam families 21

4.1 Example of SD sequences 29
4.2 Species List with Optimal Growth Temperature 33

6.1 Definition of computational Biology and Bioinformatics 61

A.1 List of Abbreviations 141

C List of Figures

3.1 Central Dogma of molecular biology 11
3.2 Chemical structure RNA 12
3.3 RNA base pairs 14
3.4 RNA secondary structure loop types 15
3.5 Secondary structure of a tRNA 19

4.1 Polycistronic mRNA 24
4.2 Accessibility of the UTR and the CDS 32
4.3 Optimal Growth and RNA Accessibility 34
4.4 Positional accessibility mRNA 35
4.5 Promoter sequence elements 40
4.6 Secondary structure of E. coli's 6S RNA 41
4.7 Scheme of Translation Initiation 44

5.1 Experimentally confirmed sRNA 50
5.2 sRNA regulation mechanism 52
5.3 Molecular structure of Hfq 53

6.1 Dot-plot representing the structure ensemble 64
6.2 Reconstructing the partition function contributions 69
6.3 Reconstructing the partition function for alternative structures 69
6.4 Dot-plot representing the local pairing probability 71
6.5 Consensus structure for alignment-folding 73

10.1 Analysis work-flow for sRNA discovery 132

D Bibliography

[1] Adhin et al. Scanning model for translational reinitiation in eubacteria. *Journal of molecular biology*, 213(4):811–818, 1990.

[2] Alkan et al. RNA–RNA interaction prediction and antisense RNA target search. *Journal of Computational Biology*, 13(2):267–282, 2006.

[3] Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[4] Andrade et al. Small RNA modules confer different stabilities and interact differently with multiple targets. *PloS one*, 8(1):e52866, 2013.

[5] Andronescu et al. RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Research*, 31(13):3416–3422, 2003.

[6] Andronescu et al. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):i19–i28, 2007.

[7] Argaman et al. fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *Journal of molecular biology*, 300(5):1101–1112, 2000.

[8] Arnold et al. Observation of Escherichia coli ribosomal proteins and their posttranslational modifications by mass spectrometry. *Analytical biochemistry*, 269(1):105–112, 1999.

[9] Babak et al. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC bioinformatics*, 8(1):33, 2007.

[10] Backofen et al. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol*, 7(1):33–42, 2010.

[11] Balleza et al. Regulation by transcription factors in bacteria: beyond description. *FEMS microbiology reviews*, 33(1):133–151, 2008.

[12] Barrick et al. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology*, 8(11):R239, 2007.

[13] Beadle et al. Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):499, 1941.

[14] Beisel et al. Multiple factors dictate target selection by Hfq-binding small RNAs. *The EMBO journal*, 31(8):1961–1974, 2012.

[15] Bejerano-Sagie et al. The role of small RNAs in quorum sensing. *Current opinion in microbiology*, 10(2):189–198, 2007.

[16] Belasco. All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nature Reviews Molecular Cell Biology*, 11(7):467–478, 2010.

[17] Bernhart et al. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, 2006.

[18] Bernhart et al. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology*, 1(1):3, 2006.

[19] Bernhart et al. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics*, 9(1):474, 2008.

[20] Bernstein et al. Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2758–2763, 2004.

[21] Binns et al. Expression of the Escherichia coli pcnB gene is translationally limited using an inefficient start codon: a second chromosomal example of translation initiated at AUU. *Molecular microbiology*, 44(5):1287–1298, 2002.

[22] Boehm et al. The csgD mRNA as a hub for signal integration via multiple small RNAs. *Molecular microbiology*, 84(1):1–5, 2012.

[23] Boisset et al. Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes & development*, 21(11):1353–1366, 2007.

[24] Boni et al. Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic acids research*, 19(1):155–162, 1991.

[25] Bonnefond et al. Exploiting protein engineering and crystal polymorphism for successful X-ray structure determination. *Crystal Growth & Design*, 2011.

[26] Brandt et al. The native 3D organization of bacterial polysomes. *Cell*, 136(2):261–271, 2009.

[27] Brantl et al. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Current opinion in microbiology*, 10(2):102–109, 2007.

[28] Brion et al. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

[29] Bronikowski et al. Evolutionary adaptation to temperature. VIII. Effects of temperature on growth rate in natural isolates of Escherichia coli and Salmonella enterica from different thermal environments. *Evolution*, 55(1):33–40, 2001.

[30] Buck et al. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, 2004.

[31] Burge et al. Rfam 11.0: 10 years of RNA families. *Nucleic acids research*, 41(D1):D226–D232, 2013.

[32] Busch et al. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 2008.

[33] Cadillo-Quiroz et al. Contribution of transcriptomics to systems-level understanding of methanogenic archaea. *Archaea*, 2013, 2013.

[34] Caldelari et al. RNA–mediated regulation in pathogenic bacteria. *Cold Spring Harbor perspectives in medicine*, 3(9):a010298, 2013.

[35] Cech et al. The ribosome is a ribozyme. *Science*, 289(5481):878–879, 2000.

[36] Chabelskaya et al. A staphylococcus aureus small RNA is required for bacterial virulence and regulates the expression of an immune-evasion molecule. *PLoS pathogens*, 6(6):e1000927, 2010.

[37] Chen et al. Determination of the optimal aligned spacing between the Shine–Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic acids research*, 22(23):4953–4957, 1994.

[38] Chen et al. Computational identification of small RNAs in Clostridium Acetobutylicum and prediction of mRNA targets. In *The 2008 Annual Meeting*, 2008.

[39] Chen et al. Small RNAs in the genus clostridium. *MBio*, 2(1), 2011.

[40] Chevalier et al. Probing mRNA structure and sRNA–mRNA interactions in bacteria using enzymes and lead (II). In *Riboswitches*, pages 215–232. Springer, 2009.

[41] Chibani-Chennoufi et al. Phage-host interaction: an ecological perspective. *Journal of bacteriology*, 186(12):3677–3686, 2004.

[42] Chitsaz et al. biRNA: Fast RNA-RNA binding sites prediction. In *Algorithms in Bioinformatics*, pages 25–36. Springer, 2009.

[43] Crick et al. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[44] Crick et al. Codon—anticodon pairing: the wobble hypothesis. *Journal of molecular biology*, 19(2):548–555, 1966.

[45] Crick et al. The genetic code – yesterday, today, and tomorrow. In *Cold Spring Harbor symposia on quantitative biology*, volume 31, pages 3–9. Cold Spring Harbor Laboratory Press, 1966.

[46] Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[47] Dahlquist et al. Interaction of translation initiation factor IF1 with the E. coli ribosomal A site. *Journal of molecular biology*, 299(1):1–15, 2000.

[48] Dalli et al. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, 22(13):1593–1599, 2006.

[49] de Avila e Silva et al. BacPP: Bacterial promoter prediction – a tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of theoretical biology*, 287:92–99, 2011.

[50] de Smit et al. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proceedings of the National Academy of Sciences*, 87(19):7668–7672, 1990.

[51] Pereira de Souza et al. Spontaneous crowding of ribosomes and proteins inside vesicles: A possible mechanism for the origin of cell metabolism. *ChemBioChem*, 12(15):2325–2330, 2011.

[52] Deigan et al. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 106(1):97–102, 2009.

[53] Dieterich et al. Computational biology of RNA interactions. *Wiley Interdisciplinary Reviews: RNA*, 4(1):107–120, 2013.

[54] Dima et al. Extracting stacking interaction parameters for RNA from the data set of native structures. *Journal of molecular biology*, 347(1):53–69, 2005.

[55] Ding et al. RNA secondary structure prediction by centroids in a boltzmann weighted ensemble. *RNA*, 11(8):1157–1166, 2005.

[56] Doetsch et al. Study of E. coli Hfq's RNA annealing acceleration and duplex destabilization activities using substrates with different GC-contents. *Nucleic acids research*, 41(1):487–497, 2013.

[57] Doshi et al. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics*, 5(1):105, 2004.

[58] Dugar et al. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple campylobacter jejuni isolates. *PLoS genetics*, 9(5):e1003495, 2013.

[59] Edgell et al. Barriers to intron promiscuity in bacteria. *Journal of Bacteriology*, 182(19):5281–5289, 2000.

[60] Fickett et al. Eukaryotic promoter recognition. *Genome Research*, 7(9):861–878, 1997.

[61] Fire et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *nature*, 391(6669):806–811, 1998.

[62] Flamm et al. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, 2000.

[63] Freier et al. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences*, 83(24):9373–9377, 1986.

[64] Fröhlich et al. Activation of gene expression by small RNA. *Current opinion in microbiology*, 12(6):674–682, 2009.

[65] Frohman. On beyond classic RACE (rapid amplification of cDNA ends). *Genome Research*, 4(1):S40–S58, 1994.

[66] Gama-Castro et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research*, 36(suppl 1):D120–D124, 2008.

[67] Gama-Castro et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research*, 39(suppl 1):D98–D105, 2011.

[68] Garrett et al. *Biochemistry.* Brooks/Cole Publishing Company, 2012.

[69] Gerdes et al. Antisense RNA-regulated programmed cell death. *Annual review of genetics*, 31(1):1–31, 1997.

[70] Gilbert. The RNA world. *Nature*, 319, 1986.

[71] Goodrich et al. From bacteria to humans, chromatin to elongation, and activation to repression: The expanding roles of noncoding RNAs in regulating transcription. *Critical reviews in biochemistry and molecular biology*, 44(1):3–15, 2009.

[72] Gordon et al. Improved prediction of bacterial transcription start sites. *Bioinformatics*, 22(2):142–148, 2006.

[73] Gottesman et al. The small RNA regulators of Escherichia coli: roles and mechanisms. *Annu. Rev. Microbiol.*, 58:303–328, 2004.

[74] Grate et al. RNA modeling using Gibbs sampling and stochastic context free grammars. In *Ismb*, volume 2, pages 138–146, 1994.

[75] Gray et al. Primary sequence of the alpha-toxin gene from Staphylococcus aureus wood 46. *Infection and immunity*, 46(2):615–618, 1984.

[76] Gruber et al. The vienna RNA websuite. *Nucleic acids research*, 36(suppl 2):W70–W74, 2008.

[77] Gruber et al. RNAz 2.0: improved noncoding RNA detection. In *Pacific Symposium on Biocomputing*, volume 15, pages 69–79. World Scientific, 2010.

[78] Gualerzi et al. Initiation of mRNA translation in prokaryotes. *Biochemistry*, 29(25):5881–5889, 1990.

[79] Guillier et al. Remodelling of the Escherichia coli outer membrane by two small regulatory RNAs. *Molecular microbiology*, 59(1):231–247, 2006.

[80] Hajiaghayi et al. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC bioinformatics*, 13(1):22, 2012.

[81] Hajnsdorf et al. Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie*, 2012.

[82] Hall et al. A role for mRNA secondary structure in the control of translation initiation. 1982.

[83] Haller et al. The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Accounts of Chemical Research*, 44(12):1339–1348, 2011.

[84] Hämmerle et al. Structural and biochemical studies on ATP binding and hydrolysis by the Escherichia coli RNA chaperone Hfq. *PloS one*, 7(11):e50892, 2012.

[85] Harada et al. Chromatin affinity purification. In *Transcriptional Regulation*, pages 237–253. Springer, 2012.

[86] Hardt et al. Role of the D arm and the anticodon arm in tRNA recognition by eubacterial and eukaryotic RNase P enzymes. *Biochemistry*, 32(48):13046–13053, 1993.

[87] He et al. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522–531, 2004.

[88] Hengge-Aronis et al. Recent insights into the general stress response regulatory network in Escherichia coli. *Journal of molecular microbiology and biotechnology*, 4(3):341–346, 2002.

[89] Henikoff et al. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[90] Herbig et al. Automated transcription start site prediction for comparative transcriptomics using the SuperGenome. *EMBnet. journal*, 19(A):pp–19, 2013.

[91] Hofacker et al. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[92] Hofacker et al. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88(1):207–237, 1998.

[93] Hofacker et al. Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology*, 319(5):1059–1066, 2002.

[94] Horvath et al. CRISPR/Cas, the immune system of bacteria and archaea. *Science*, 327(5962):167–170, 2010.

[95] Howden et al. Analysis of the small RNA transcriptional response in multidrug resistant Staphylococcus aureus after antimicrobial exposure. *Antimicrobial agents and chemotherapy*, 2013.

[96] Huang et al. sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic acids research*, 37(suppl 1):D150–D154, 2009.

[97] Huang et al. Target prediction and a statistical sampling algorithm for RNA–RNA interaction. *Bioinformatics*, 26(2):175–181, 2010.

[98] Huerta et al. NIH working definition of bioinformatics and computational biology. *US National Institute of Health*, 2000.

[99] Huntzinger et al. Staphylococcus aureus RNAIII and the endoribonuclease III coordinately regulate spa gene expression. *The EMBO journal*, 24(4):824–835, 2005.

[100] Iserentant et al. Secondary structure of mRNA and efficiency of translation initiation. *Gene*, 9(1):1–12, 1980.

[101] Jiang et al. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, 9(1):192, 2008.

[102] Jin. Influences on gene expression in vivo by a Shine–Dalgarno sequence. *Molecular microbiology*, 60(2):480–492, 2006.

[103] Jin et al. Small noncoding RNA GcvB is a novel regulator of acid resistance in escherichia coli. *BMC genomics*, 10(1):165, 2009.

[104] Jones et al. In vivo translational start site selection on leaderless mRNA transcribed from the Streptomyces fradiae aph gene. *Journal of bacteriology*, 174(14):4753–4760, 1992.

[105] Joseph et al. Regulation of galactose operon expression: glucose effects and role of cyclic adenosine 3', 5'-monophosphate. *Journal of bacteriology*, 146(1):149–154, 1981.

[106] Kajitani et al. Regulation of the Escherichia coli hfq gene encoding the host factor for phage Q beta. *Journal of bacteriology*, 176(2):531–534, 1994.

[107] Kanhere et al. A novel method for prokaryotic promoter prediction based on DNA stability. *BMC bioinformatics*, 6(1):1, 2005.

[108] Kanhere et al. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic acids research*, 33(10):3165–3175, 2005.

[109] Kastelein et al. Opening the closed ribosome-binding site of the lysis cistron of bacteriophage MS2. 1983.

[110] Katz et al. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Research*, 13(9):2042–2051, 2003.

[111] Kawamoto et al. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Molecular microbiology*, 61(4):1013–1022, 2006.

[112] Keseler et al. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic acids research*, 39(suppl 1):D583–D590, 2011.

[113] Kingsford et al. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome biology*, 8(2):R22, 2007.

[114] Klein et al. RSEARCH: finding homologs of single structured RNA sequences. *Bmc Bioinformatics*, 4(1):44, 2003.

[115] Klipcan et al. Optimal growth temperature of prokaryotes correlates with class II amino acid composition. *FEBS letters*, 580(6):1672–1676, 2006.

[116] Komarova et al. AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in Escherichia coli. *Journal of bacteriology*, 187(4):1344–1349, 2005.

[117] Kornblihtt et al. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology*, 2013.

[118] Kortmann et al. Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews Microbiology*, 10(4):255–265, 2012.

[119] Koshland. The seven pillars of life. *Science*, 295(5563):2215–2216, 2002.

[120] Krebs et al. *Lewin's genes X*. Jones & Bartlett Learning, 2009.

[121] Kruger et al. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *cell*, 31(1):147–157, 1982.

[122] Kudla et al. Coding-sequence determinants of gene expression in Escherichia coli. *science*, 324(5924):255–258, 2009.

[123] Latif et al. The genome organization of Thermotoga maritima reflects its lifestyle. *PLoS genetics*, 9(4):e1003485, 2013.

[124] Laursen et al. Initiation of protein synthesis in bacteria. *Microbiology and molecular biology reviews*, 69(1):101–123, 2005.

[125] Leontis et al. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

[126] Li et al. Predicting sRNAs and their targets in bacteria. *Genomics, proteomics & bioinformatics*, 2012.

[127] Link et al. Structure of escherichia coli Hfq bound to polyriboadenylate RNA. *Proceedings of the National Academy of Sciences*, 106(46):19292–19297, 2009.

[128] Livny et al. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic acids research*, 33(13):4096–4105, 2005.

[129] Livny et al. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PloS one*, 3(9):e3197, 2008.

[130] Lobry et al. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene*, 385(1):128–136, 2006.

[131] Lorenz et al. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[132] Lorenz et al. RNA folding algorithms with G-quadruplexes. In *Advances in Bioinformatics and Computational Biology*, pages 49–60. Springer, 2012.

[133] Lorenz et al. 2D meets 4G: G-quadruplexes in RNA secondary structure prediction. 2013.

[134] Low et al. SHAPE-directed RNA secondary structure prediction. *Methods*, 52(2):150–158, 2010.

[135] Majdalani et al. Bacterial small RNA regulators. *Critical reviews in biochemistry and molecular biology*, 40(2):93–113, 2005.

[136] Makarova et al. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct*, 1(1):7, 2006.

[137] Mann. Phages of the marine cyanobacterial picophytoplankton. *FEMS microbiology reviews*, 27(1):17–34, 2003.

[138] Mariner et al. Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Molecular cell*, 29(4):499–509, 2008.

[139] Martin-Farmer et al. A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in Escherichia coli. *Molecular microbiology*, 31(4):1025–1038, 1999.

[140] Massé et al. Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli. *Genes & development*, 17(19):2374–2383, 2003.

[141] Massé et al. Effect of RyhB small RNA on global iron use in Escherichia coli. *Journal of bacteriology*, 187(20):6962–6971, 2005.

[142] Massé et al. Small RNAs controlling iron metabolism. *Current opinion in microbiology*, 10(2):140–145, 2007.

[143] McCarthy et al. Prokaryotic translation: the interactive pathway leading to initiation. *Trends in Genetics*, 10(11):402–407, 1994.

[144] McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[145] McCown et al. An expanded collection and refined consensus model of glmS ribozymes. *RNA*, 17(4):728–736, 2011.

[146] McEntyre et al. The NCBI handbook. 2002.

[147] McKay. What is life – and how do we search for it in other worlds? *PLoS biology*, 2(9):e302, 2004.

[148] Mehta et al. A quantitative comparison of sRNA-based and protein-based gene regulation. *Molecular systems biology*, 4(1), 2008.

[149] Mendoza-Vargas et al. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One*, 4(10):e7526, 2009.

[150] Mentz et al. Comprehensive discovery and characterization of small RNAs in corynebacterium glutamicum ATCC 13032. *BMC genomics*, 14(1):714, 2013.

[151] Meyer et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research*, 41(D1):D64–D69, 2013.

[152] Miranda-Ríos et al. A conserved RNA structure thi box is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proceedings of the National Academy of Sciences*, 98(17):9736–9741, 2001.

[153] Mizuno et al. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proceedings of the National Academy of Sciences*, 81(7):1966, 1984.

[154] Moll et al. Evidence against an interaction between the mRNA downstream box and 16S rRNA in translation initiation. *Journal of bacteriology*, 183(11):3499–3505, 2001.

[155] Moll et al. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Molecular microbiology*, 43(1):239–246, 2002.

[156] Møller et al. Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes & development*, 16(13):1696–1706, 2002.

[157] Mondragón et al. Structural studies of RNase P. *Annual review of biophysics*, 42:537–557, 2013.

[158] Morfeldt et al. Activation of alpha-toxin translation in Staphylococcus aureus by the trans-encoded antisense RNA, RNAIII. *The EMBO journal*, 14(18):4569, 1995.

[159] Mückstein et al. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.

[160] Mückstein et al. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. In *Bioinformatics research and development*, pages 114–127. Springer, 2008.

[161] Nawrocki et al. Computational identification of functional RNA homologs in metagenomic data. *RNA biology*, 10(7):1–10, 2013.

[162] Nicholson. Function, mechanism and regulation of bacterial ribonucleases. *FEMS microbiology reviews*, 23(3):371–390, 1999.

[163] Nussinov et al. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82, 1978.

[164] The Students of the Bioinformatics II Lab Class 2013 et al. The trouble with long-range base pairs in RNA folding. *Advances in Bioinformatics and Computational Biology*, pages 1–11, 2013.

[165] Olivarius et al. High-throughput verification of transcriptional starting sites by deep-RACE. *BioTechniques*, 46(2):130, 2009.

[166] Opdyke et al. GadY, a small-RNA regulator of acid response genes in Escherichia coli. *Journal of bacteriology*, 186(20):6698–6705, 2004.

[167] Osterman et al. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic acids research*, 41(1):474–486, 2013.

[168] Ott et al. NAPP: the nucleic acid phylogenetic profile database. *Nucleic acids research*, 40(D1):D205–D209, 2012.

[169] Pannone et al. RNA degradation: Sm-like proteins wRING the neck of mRNA. *Current Biology*, 10(13):R478–R481, 2000.

[170] Papenfort et al. Evidence for an autonomous 5' target recognition domain in an Hfq-associated small RNA. *Proceedings of the National Academy of Sciences*, 107(47):20435–20440, 2010.

[171] Pedersen et al. The biology of eukaryotic promoter prediction – a review. *Computers & Chemistry*, 23(3):191–207, 1999.

[172] Peters et al. Rho directs widespread termination of intragenic and stable RNA transcription. *Proceedings of the National Academy of Sciences*, 106(36):15406–15411, 2009.

[173] Peters et al. Bacterial transcription terminators: the RNA 3' end chronicles. *Journal of molecular biology*, 412(5):793–813, 2011.

[174] Petrelli et al. Translation initiation factor IF3: two domains, five functions, one mechanism? *The EMBO journal*, 20(16):4560–4569, 2001.

[175] Pfeiffer et al. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nature structural & molecular biology*, 16(8):840–846, 2009.

[176] Pipas et al. Method for predicting RNA secondary structure. *Proceedings of the National Academy of Sciences*, 72(6):2017–2021, 1975.

[177] Pleij et al. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Research*, 13(5):1717–1731, 1985.

[178] Prévost et al. Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage. *Genes & development*, 25(4):385–396, 2011.

[179] Queen et al. Differential translation efficiency explains discoordinate expression of the galactose operon. *Cell*, 25(1):241–249, 1981.

[180] Raghavan et al. Genome-wide detection of novel regulatory RNAs in E. coli. *Genome Research*, 21(9):1487–1497, 2011.

[181] Ramachandran et al. The architecture and ppGpp-dependent expression of the primary transcriptome of Salmonella Typhimurium during invasion gene expression. *BMC genomics*, 13(1):25, 2012.

[182] Ramadoss et al. tmRNA is essential in Shigella flexneri. *PloS one*, 8(2):e57537, 2013.

[183] Rangannan et al. High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics*, 26(24):3043–3050, 2010.

[184] Rehmsmeier et al. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[185] Ringquist et al. Translation initiation in Escherichia coli: sequences within the ribosome-binding site. *Molecular microbiology*, 6(9):1219–1229, 1992.

[186] Rinn et al. Functional demarcation of active and silent chromatin domains in human HOX loci by non-coding RNAs. *Cell*, 129(7):1311, 2007.

[187] Rivas et al. The four ingredients of single-sequence RNA secondary structure prediction: A unifying perspective. *RNA biology*, 10(7), 2013.

[188] Rodnina et al. The ribosome as a versatile catalyst: reactions at the peptidyl transferase center. *Current opinion in structural biology*, 2013.

[189] Rodrigo et al. A new frontier in synthetic biology: automated design of small RNA devices in bacteria. *Trends in Genetics*, 29(9):529–536, 2013.

[190] Romeo et al. Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Molecular microbiology*, 29(6):1321–1330, 1998.

[191] Salari et al. Fast prediction of RNA-RNA interaction. *Algorithms for molecular Biology*, 5(5), 2010.

[192] Salgado et al. Operons in Escherichia coli: genomic analyses and predictions. *Proceedings of the National Academy of Sciences*, 97(12):6652–6657, 2000.

[193] Salim et al. An upstream Hfq binding site in the fhlA mRNA leader region facilitates the OxyS-fhlA interaction. *PLoS One*, 5(9):e13028, 2010.

[194] Schmidtke et al. Genome-wide transcriptome analysis of the plant pathogen xanthomonas identifies sRNAs with putative virulence functions. *Nucleic acids research*, 40(5):2020–2031, 2012.

[195] Schumacher et al. Structures of the pleiotropic translational regulator Hfq and an Hfq–RNA complex: a bacterial Sm-like protein. *The EMBO journal*, 21(13):3546–3556, 2002.

[196] Scolnick et al. Release factors differing in specificity for terminator codons. *Proceedings of the National Academy of Sciences of the United States of America*, 61(2):768, 1968.

[197] Shabalina et al. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic acids research*, 34(8):2428–2437, 2006.

[198] Shamovsky et al. RNA-mediated response to heat shock in mammalian cells. *Nature*, 440(7083):556–560, 2006.

[199] Sharma et al. Experimental approaches for the discovery and characterization of regulatory small RNA. *Current opinion in microbiology*, 12(5):536–546, 2009.

[200] Sharma et al. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286):250–255, 2010.

[201] Sharp et al. Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research*, 33(4):1141–1153, 2005.

[202] Shean et al. Translation of the prophage λ cl transcript. *Cell*, 70(3):513–522, 1992.

[203] Shinhara et al. Deep sequencing reveals as-yet-undiscovered small RNAs in Escherichia coli. *BMC genomics*, 12(1):428, 2011.

[204] Shiraki et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, 2003.

[205] Siebert et al. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–3359, 2005.

[206] Sittka et al. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS genetics*, 4(8):e1000163, 2008.

[207] Sledjeski et al. A small RNA acts as an antisilencer of the H-NS-silenced rcsA gene of Escherichia coli. *Proceedings of the National Academy of Sciences*, 92(6):2003–2007, 1995.

[208] Smith et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[209] Soutourina et al. Genome-wide identification of regulatory RNAs in the human pathogen clostridium difficile. *PLoS genetics*, 9(5):e1003493, 2013.

[210] Šponer et al. Nature and magnitude of aromatic base stacking in DNA and RNA: Quantum chemistry, molecular mechanics and experiment. *Biopolymers*, 2013.

[211] Sprengart et al. The initiation of translation in E. coli: apparent base pairing between the 16S rRNA and downstream sequences of the mRNA. *Nucleic acids research*, 18(7):1719–1723, 1990.

[212] Sprengart et al. The downstream box: an efficient and independent translation initiation signal in Escherichia coli. *The EMBO journal*, 15(3):665, 1996.

[213] Springer et al. RNA mimicry in the translational apparatus. *Cold Spring Harbor Monograph Archive*, 35:377–413, 1998.

[214] Staple et al. Pseudoknots: RNA structures with diverse functions. *PLoS biology*, 3(6):e213, 2005.

[215] Stark et al. Ribonuclease P: an enzyme with an essential RNA component. *Proceedings of the National Academy of Sciences*, 75(8):3717–3721, 1978.

[216] Stevens et al. In silico estimation of translation efficiency in human cell lines: Potential evidence for widespread translational control. *PloS one*, 8(2):e57625, 2013.

[217] Studnicka et al. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Research*, 5(9):3365–3388, 1978.

[218] Sun. Characterization of the small RNA transcriptome in plant–microbe (brassica/erwinia) interactions by high-throughput sequencing. *Biotechnology letters*, pages 1–11, 2013.

[219] Tafer et al. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*, 24(22):2657–2663, 2008.

[220] Taylor et al. From MicrobeWiki, the student-edited microbiology resource.

[221] Temlyakova et al. 70 electrostatic properties of bacterial DNA and promoter predictions. *Journal of Biomolecular Structure and Dynamics*, 31(sup1):44–45, 2013.

[222] Thompson et al. Characterization of the 5'-terminal structure of simian virus 40 early mRNA's. *Journal of virology*, 31(2):437–446, 1979.

[223] Tjaden. TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic acids research*, 36(suppl 2):W109–W113, 2008.

[224] Toledo-Arana et al. Small noncoding RNAs controlling pathogenesis. *Current opinion in microbiology*, 10(2):182–188, 2007.

[225] Toledo-Arana et al. The listeria transcriptional landscape from saprophytism to virulence. *Nature*, 459(7249):950–956, 2009.

[226] Trotochaud et al. 6S RNA function enhances long-term cell survival. *Journal of bacteriology*, 186(15):4978–4985, 2004.

[227] Tzareva et al. Ribosome-messenger recognition in the absence of the Shine–Dalgarno interactions. *FEBS letters*, 337(2):189–194, 1994.

[228] Ullmann et al. Cyclic AMP as a modulator of polarity in polycistronic transcriptional units. *Proceedings of the National Academy of Sciences*, 76(7):3194–3197, 1979.

[229] Ulveling et al. When one is better than two: RNA with dual functions. *Biochimie*, 93(4):633–644, 2011.

[230] Urban et al. A conserved small RNA promotes discoordinate expression of the glmUS operon mRNA to activate GlmS synthesis. *Journal of molecular biology*, 373(3):521–528, 2007.

[231] Urban et al. Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic acids research*, 35(3):1018–1037, 2007.

[232] Valentin-Hansen et al. MicroReview: The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Molecular microbiology*, 51(6):1525–1533, 2004.

[233] Valle et al. Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. *Nature Structural & Molecular Biology*, 10(11):899–906, 2003.

[234] Večerek et al. Interaction of the RNA chaperone Hfq with mRNAs: direct and indirect roles of Hfq in iron metabolism of Escherichia coli. *Molecular microbiology*, 50(3):897–909, 2003.

[235] Vesper et al. Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in Escherichia coli. *Cell*, 147(1):147–157, 2011.

[236] Vockenhuber et al. Streptomyces coelicolor sRNA scr5239 inhibits agarase expression by direct base pairing to the dagA coding region. *Microbiology*, 158(2):424–435, 2012.

[237] Vogel et al. How to find small non-coding RNAs in bacteria. *Biological chemistry*, 386(12):1219–1238, 2005.

[238] Voß et al. Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC genomics*, 10(1):123, 2009.

[239] Wachsmuth et al. De novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic acids research*, 41(4):2541–2551, 2013.

[240] Wadler et al. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proceedings of the National Academy of Sciences*, 104(51):20454–20459, 2007.

[241] Warner et al. The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, 24(11):437–440, 1999.

[242] Washietl et al. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–2459, 2005.

[243] Washietl et al. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic acids research*, 40(10):4261–4272, 2012.

[244] Waters et al. Regulatory RNAs in bacteria. *Cell*, 136(4):615–628, 2009.

[245] Wenzel et al. RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, 28(21):2738–2746, 2012.

[246] Witkos et al. Practical aspects of microRNA target prediction. *Current molecular medicine*, 11(2):93, 2011.

[247] Wood et al. The influence of messenger RNA secondary structure on expression of an immunoglobulin heavy chain in Escherichia coli. *Nucleic acids research*, 12(9):3937–3950, 1984.

[248] Wootton et al. Bacterial pathogenesis: sRNA clears the way for G4. *Nature Reviews Microbiology*, 11(3):146–147, 2013.

[249] Workman et al. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Research*, 27(24):4816–4822, 1999.

[250] Wurtzel et al. A single-base resolution map of an archaeal transcriptome. *Genome research*, 20(1):133–141, 2010.

[251] Wurtzel et al. Comparative transcriptomics of pathogenic and non-pathogenic listeria species. *Molecular systems biology*, 8(1), 2012.

[252] Yakhnin et al. CsrA represses translation of sdiA, which encodes the N-acylhomoserine-L-lactone receptor of Escherichia coli, by binding exclusively within the coding region of sdiA mRNA. *Journal of bacteriology*, 193(22):6162–6170, 2011.

[253] Yan et al. Determination of sRNA expressions by RNA-seq in yersinia pestis grown in vitro and during infection. *PloS one*, 8(9):e74495, 2013.

[254] Zhang et al. Global analysis of small RNA and mRNA targets of Hfq. *Molecular microbiology*, 50(4):1111–1124, 2003.

[255] Zhou et al. Crystal structures of 70S ribosomes bound to release factors RF1, RF2 and RF3. *Current Opinion in Structural Biology*, 2012.

[256] Höner zu Siederdissen. *Grammatical Approaches to Problems in RNA Bioinformatics*. PhD thesis, University Vienna, 2013.

[257] Zuker et al. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.

E Curriculum Vitae

Fabian Amman

Born on the 23^{rd} of January 1981 in Feldkirch, Austria, to Lydia and Werner Amman.

EDUCATION

Studies in *Astronomy*, University Vienna, 1999-2000;
Studies in *Genetics and Microbiology*, University Vienna, 2000-2007;
Diploma thesis in the Group of Prof. Dr. Christopher Gerner at the General Hospital Vienna (AKH, Institute for Cancer Research) on the usage of histone isotypes in cell proliferation and differentiation, 2006-2007;
Studies in *Agricultural Engineering & Water-resource Management*, University of Natural Resources and Life Sciences (BOKU), Vienna, 2010-2012;
Doctoral research study in the Group of Prof. Dr. Ivo Hofacker at the University Vienna (Institute for theoretical biochemistry) on small non-coding RNA in bacteria, 2009-2013;

WORK EXPERIENCE

Internship in the *Food-Safety and QC laboratory* at Hilcona AG in Schaan/Liechtenstein, Summer 2002;
Internship in the Group of Prof. Dr. Christian Schlötterer, Institute for *Animal Breeding and Genetics*, University of Veterinary Medicine, Vienna, Winter 2004/2005;
Alternative civilian service in the Hôpital Ngaoubela/Cameroon; Employed in the Hospital laboratory and responsible for the design and management of a *Mother-Child-Health project*, 2008-2009;
Researcher in the Group of Prof. Dr. Peter Stadler, Interdisciplinary Center for Bioinformatics, University Leipzig, since 2013;

Fast accessibility-based prediction of RNA–RNA interactions

Hakim Tafer, Fabian Amman, Florian Eggenhofer, Peter Stadler, and Ivo Hofacker
Bioinformatics (27/14/1934–1940);
2011
doi:10.1093/bioinformatics/btr281

Animal snoRNAs and scaRNAs with exceptional structures

Manja Marz, Andreas Gruber, Christian Höner zu Siederdissen, Fabian Amman, Stefan Badelt, Sebastian Bartschat, Stephan Bernhart, Wolfgang Beyer, Stephanie Kehr, Ronny Lorenz, Andrea Tanzer, Dilmurat Yusuf, Hakim Tafer, Ivo Hofacker, and Peter Stadler
RNA Biology (8/6/938–946);
2011
doi:10.4161/rna.8.6.16603

RNA folding algorithms with G-quadruplexes

Ronny Lorenz, Stephan Bernhart, Fabian Externbrink, Jing Qin, Christian Höner zu Siederdissen, Fabian Amman, Ivo Hofacker, and Peter Stadler
Advances in Bioinformatics and Computational Biology (7409/-/49–60);
2012
doi:10.1007/978-3-642-31927-3

Modelling Translation Initiation under the Influence of sRNA

Fabian Amman, Christoph Flamm, and Ivo Hofacker
International journal of molecular sciences (13/12/16223–16240);
2012
doi:10.3390/ijms131216223

2D meets 4G: G-quadruplexes in RNA secondary structure prediction

Ronny Lorenz, Stephan Bernhart, Jing Qin, Andrea Tanzer, Fabian Amman, Ivo Hofacker, and Peter Stadler
IEEE (-/-/-);
2013
doi:10.1109/tcbb.2013.7

Alterations of the Transcriptome of Sulfolobus acidocaldarius by Exoribonuclease aCPSF2

Birgit Märtens, Fabian Amman, Salim Manoharadas, Lukas Zeichen, Alvaro Orell, Sonja-Verena Albers, Ivo Hofacker, and Udo Bläsi
PloS one; (8/10/e76569);
2013
DOI:10.1371/JOURNAL.PONE.0076569

The Trouble with Long-Range Base Pairs in RNA Folding

The Students of the Bioinformatics II Lab Class 2013, Fabian Amman, Stephan Bernhart, Gero Doose, Ivo Hofacker, Jing Qin, Peter F. Stadler, and Sebastian Will
Advances in Bioinformatics and Computational Biology; (*8213/-/1–11*);
2013
DOI:10.1007/978-3-319-02624-4

Archaeal Signal Transduction: Impact of Protein Phosphatase Deletions on Cell Size, Motility and Energy Metabolism in Sulfolobus acidocaldarius

Julia Reimann, Dominik Esser, Alvaro Orell, Fabian Amman, Trong Khoa Pham, Josselin Noirel, Ann-Christin Lindas, Rolf Bernander, Phillip Wright, Bettina Siebers, and Sonja-Verena Albers
Molecular & Cellular Proteomics; *accepted (September 27th, 2013)*

TSSAR: TSS Annotation Regime for dRNA-seq data

Fabian Amman, Michael T. Wolfinger, Ronny Lorenz, Ivo Hofacker, Peter Stadler, and Sven Findeiß
BMC Bioinformatics; *submitted (June 19th, 2013)*