



universität
wien

DIPLOMARBEIT

Titel der Diplomarbeit

RNApredator: A web-based tool to predict small RNA
targets

Verfasser

Florian Eggenhofer

angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, 2011

Studienkennzahl lt. Studienblatt:

A 490

Studienrichtung lt. Studienblatt:

Diplomstudium Molekulare Biologie

Betreuer:

Univ.-Prof. Dipl.-Phys. Dr. Ivo L. Hofacker

Danksagung

Meiner Mutter und meinem Vater für die Möglichkeit zu Studieren.

Ivo für Inspiration, Geduld, Enthusiasmus und das Privileg bei dir Diplomarbeit zu schreiben.

Hakim, Fabian und Peter für die gute Zusammenarbeit und fruchtbare Diskussionen.

Bernie, Christian und Sven, für Korrekturlesen und viele hilfreiche Tipps.

Jörg, dafür das er ein humorvoller, kompetenter und verlässlicher Zimmerkollege und Freund ist.

Judith und Richard für administrative Unterstützung.

Meinen Arbeitskollegen am TBI für eine freundliche und gesellige Atmosphäre, in die man immer wieder gerne zurück kehrt.

Abstract

A multitude of small non-coding RNAs (sRNAs) is encoded by bacterial genomes. These sRNAs are heterogeneous in structure, function and size. The majority of sRNAs functions as post-transcriptional regulators by means of specific hybridization with the 5-untranslated region of mRNA transcripts, thereby modifying the target transcript its ability to be translated.

At the moment about 150 sRNAs have been identified and functionally characterized, with 80 of them found in *Escherichia coli*, leaving significant potential for new isolations from other species. These will require extensive experimental analysis, which can be refined and accelerated by tools like RNApredator.

RNApredator [1], a web tool for prediction of sRNA targets, uses a dynamic programming approach (RNAplex), to compute the best putative interaction partners. A set of over 2155 genomes and plasmids from 1183 bacterial species is available for selection as target.

Compared to web servers with a similar task, RNApredator takes the accessibility of the target during the target search into account, improving the specificity of the predictions.

Additionally, enrichment in Gene Ontology terms as well as changes in accessibilities along the target sequence can be done in fully automated post-processing steps. This can provide clues about the biological function and the influence on regulation of specific mRNAs by a sRNA.

The prediction sensitivity of the underlying dynamic programming approach RNAplex is similar to that of more complex methods, but needs at least three orders of magnitude less time to complete. This makes genome-wide interaction-studies feasible.

RNApredator is available at <http://rna.tbi.univie.ac.at/RNApredator> and has been published as journal article [1].

Zusammenfassung

Bakterielle Genome codieren für eine Vielzahl von kleinen nicht-kodierenden RNAs (sRNA), welche heterogen in Struktur und Funktion, als auch in ihrer Größe sind. sRNAs besitzen regulatorische Eigenschaften, unter anderem durch Hybridisierung mit messenger RNAs (mRNAs). Dabei kann durch Bindung an den nicht codierenden 5'-Teil die Translatierbarkeit des Transkriptes verändert werden.

Aktuell sind etwa 150 sRNAs identifiziert und funktionell charakterisiert, wobei circa 80 davon aus *Escherichia coli* stammen. Dies lässt auf eine enorme Menge bis jetzt nicht isolierter sRNAs in anderen Spezies schließen. Die Analyse dieser sRNAs wird massive Anzahl von Experimenten benötigen, welche durch Werkzeuge wie RNAPredator vereinfacht und beschleunigt werden können.

RNAPredator ist ein web-basiertes Werkzeug das die Simulation solcher Wechselwirkungen zwischen sRNAs und mRNAs ermöglicht. RNAPlex welches auf einer dynamic programming Strategie basiert wird dabei zur Berechnung der besten moeglichen Interaktionspartner genutzt.

Verglichen mit anderen Web-Servern die ähnliche Funktionalität bieten, berücksichtigt RNAPredator intramolekulare Bindungen von sRNA und mRNA, was die Genauigkeit der Interaktions-Vorhersage verbessert.

Darüberhinaus, kann die Anreicherung von Gene Ontology terms und Veränderung von struktureller Zugänglichkeit bei Gruppen von ausgewählten mRNAs automatisch berechnet werden. Dies kann Anhaltspunkte für die biologische Funktion der sRNA, bzw. ihren Einfluss auf einzelne Transkripte liefern.

Die Empfindlichkeit von RNAPlex ist vergleichbar mit jener von komplexeren Methoden, benötigt aber zumindest drei Größenordnungen weniger Rechenzeit, was die Anwendung auf genomeweite Interaktionsvorhersagen durchführbar macht.

RNAPredator kann über <http://rna.tbi.univie.ac.at/RNAPredator> erreicht werden und ist als Artikel veröffentlicht worden [1].

Contents

1	Introduction	1
2	Biological Background	3
2.1	Gene expression	3
2.1.1	DNA	4
2.1.2	RNA	6
2.1.3	Protein	10
2.1.4	Transcription	12
2.1.5	Translation	16
2.2	Regulatory non-coding RNAs	24
2.2.1	ncRNAs interacting with messenger RNAs	24
2.2.2	ncRNAs interacting with proteins	27
3	Bioinformatics	29
3.1	RNA-Bioinformatics	29
3.2	RNA Folding	29
3.2.1	Additive energy model	30
3.2.2	Folding recursion	30
3.2.3	Partition function	31
3.2.4	Base pairing probability	32
3.2.5	RNAplfold	32
3.2.6	Visualization of secondary structures	33
3.3	RNA-RNA Interaction	33
3.3.1	Sequence Based Methods	34
3.3.2	RNA Cofolding	34
3.3.3	RNA Hybridization	35
3.3.4	RNAup	35
3.3.5	RNAplex	35
3.4	Gene ontology	35
3.4.1	Onthologies	36
3.4.2	GO-term	36
4	Methods	37
4.1	Data preparation	37
4.1.1	Hypothetical transcript construction	37
4.1.2	Structural accessibility calculation	38
4.2	Target prediction	39
4.3	Target evaluation	41
4.3.1	Energy-based ranking	41

4.3.2	Change in regulation	41
4.3.3	GO-term enrichment	42
5	Results - RNApredator	43
5.1	Motivation	43
5.2	Overview of RNApredator functionality	44
5.3	Pipeline and implementation	44
5.4	Input	45
5.5	Output	48
5.6	Benchmark	51
5.7	Usage statistics	52
6	Discussion and outlook	55

List of Figures

2.1	Central Dogma of Molecular Biology	3
2.2	DNA double strand	5
2.3	Heterocyclic aromatic organic compounds	5
2.4	RNA double strand	7
2.5	Nucleotide edges	8
2.6	Guanine-Uracil wobble base pair	8
2.7	RNA secondary structures	9
2.8	Amino acid monomers	11
2.9	Tripeptide	12
2.10	Schematic polycistronic transcription unit	13
2.11	Transcription	14
2.12	Schematic representation of bacterial mRNAs	17
2.13	Amino acid code	18
2.14	Translation	18
2.15	Translation initiation	19
2.16	Translation initiation region	19
2.17	Translation elongation	21
2.18	Translation termination	22
2.19	Overview of ncRNA action	24
2.20	Translational activation by increase of structural accessibility	26
2.21	Translational activation by stabilization	26
2.22	Translational inhibition by blocking functional regions	27
2.23	Translational inhibition by structural changes	27
2.24	Small RNA mediated RNA degradation	28
3.1	Examples for secondary structure visualization	34
4.1	Construction of a hypothetical transcript	37
4.2	Distribution of mRNA number per genome/plasmid	40
5.1	RNApredator pipeline visualization	45
5.2	Navigation bar of RNApredator	46
5.3	Phylogenetic tree	46
5.4	Genome search field	47
5.5	Genome search results	47
5.6	Steps required for target prediction	48
5.7	Prediction progress	48
5.8	RNApredator result list	49
5.9	Postprocessing - Interaction	50

5.10	Accessibility profiles	52
5.11	Usage-statistics of RNAPredator	54

List of Tables

3.1	Arrays used in the folding recursion	31
3.2	Folding recursion	31
5.1	Cellular component enrichment statistics	51
5.2	Benchmark of RNApredator	53

1 Introduction

In the last years the perceived role of RNA has been fundamentally transformed. What started as a rather curious in-between of DNA and Protein has changed into a subject of major importance. Early on it has been recognized that essential parts of the cellular machinery are driven by RNA-based components, for example the ribosomes. The RNA world hypothesis [2] introduced the idea that these molecules belong to the most ancient components of the cell, being a heritage of a world where RNA molecules were responsible both for storage of genetic information and for the correct cell function. In the course of evolution, RNA lost both of these functions: DNA supplemented RNA as a stabler medium of storage for genetic information in most organisms, while proteins imposed themselves as more versatile catalysts, leading to the alteration of RNA to a simple messenger quietly transporting genetic information.

Many parts of the genome, specifically the ones not coding for proteins, have been classified as "Junk DNA" [3] in the past. At the time the first sequenced genomes became available a lot of proteins were already characterized and their biological relevance well established [4, 5, 6]. It seemed clear that once the structure and function of proteins had been understood, all of the remaining mechanisms would be easily derivable from that.

The isolation and identification of a growing number of non-protein-coding RNAs transcribed from the junk DNA, bearing catalytic and/or regulatory functions, shed a new light on the importance of the junk DNA. Since that, a variety of experimental and bioinformatics-based approaches to detect and characterize them have been developed.

Layers of cellular complexity have been revealed by this, which leads directly to the motivation of this diploma-thesis. It became clear that there are RNA-mediated regulatory circuits, present in all species, which are responsible for key decisions in the cells reaction environment.

A bottom-up approach is best indicated to understand these systems as their complexity increases drastically when moving along the taxonomic tree from prokaryotic towards eukaryotic cells. Bacterial cells and their plasmids are very useful to understand the basic concepts that govern the impact of RNA on cellular regulation.

Generally small RNAs (sRNAs) influence gene expression by up-, or down-regulation of either transcription or translation. The main focus of this work

was to predict interactions of small regulatory non-coding bacterial RNAs with messenger RNAs and investigate the consequences of such processes.

At the moment 80 sRNAs [7] are annotated in *Escherichia coli* and for many of them binding partners and biological function have been analyzed, which further underlines the scientific relevance of small RNAs.

After recapitulating the background for this class of molecules and the related regulatory mechanisms, a web-based prediction tool and its program pipeline will be presented. A discussion of the tools relevance, application and an outlook will conclude this work.

2 Biological Background

To understand the overall picture an overview about gene expression is provided. The goal of this section is to establish the necessary biological background, to describe the role of small bacterial RNAs in the cell and to show the complexity of the involved biological processes and why abstraction is necessary to make these problems tractable.

This work is focused on bacteria, with most of the data originating from the *E. coli* model organism. The bacterial variant of mechanisms is described in the biological background. Archea and eukarya will be briefly discussed in the outlook and discussion section, giving a perspective where else these methods and tools applied to bacteria are relevant.

2.1 Gene expression

Our current understanding of gene expression is based on the central dogma of molecular biology [8], which in essence describes the information flow between different bio-polymers. Three principal classes of such molecules are known: DNA, RNA and Protein. Each of them consists of different sets of monomers which can be combined to form long sequences. The order of monomers within the sequence is also what is relevant for the central dogma of molecular biology.

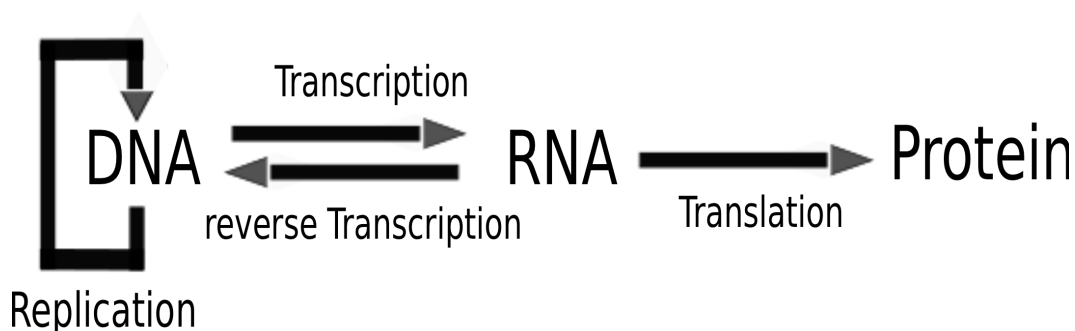


Figure 2.1: Central Dogma of Molecular Biology. The arrows denote the information flow between the molecule classes.

The classical view of gene-expression was that sequence information strictly flows from DNA to RNA(via transcription) and then to Protein(via translation), as shown in Figure 2.1. Additionally to this we are nowadays aware of the possibility of converting RNA information to DNA, by reverse transcription.

These macromolecules can then be characterized by applying four levels of structural complexity [9]:

- Primary structure:
Sequence of monomer units in a linear polymer.
- Secondary structure:
Regular local folding pattern of a polymeric molecule.
- Tertiary structure:
Complex three-dimensional shape of a folded polymer chain, especially for protein or RNA molecule.
- Quaternary structure:
Three-dimensional relationship of different macromolecules in a multi-subunit protein or protein complex.

A detailed description of the three classes follows with focus on RNA.

2.1.1 DNA

Deoxyribonucleic acid (DNA) functions as the primary information repository of all living organisms. It is duplicated by a process designated replication before each cell division.

This molecule is stable and can be repaired in case of damage. The monomers of DNA, called nucleotides consist of a nucleobase, D-deoxyribose and phosphate. They are connected with each other in a linear unbranched manner by a covalent bond between phosphate of one monomer and sugar of the next one. The carbon atoms forming the ring structure of the ribose molecule are labeled by convention as follows: 1' - connected to the nucleobase, 2' - deoxygenated, 3' - connected to the following nucleotide, 5' to the previous nucleotide (Figure 2.2). A strand of nucleotides therefore has a direction from 5' end to 3' end. DNA polymers in cells are enzymatically polymerized in the same direction.

Four nucleobases are generally found in DNA and can be differentiated in two groups based on the ring structure of organic molecules. Those groups are named Purine and Pyrimidine (see Figure 2.3) and both belong to the group of heterocyclic aromatic organic compounds, which they are derived from. Purines consist of a Pyrimidine ring fused to an Imidazole ring. Purine bases are Adenine and Guanine. Pyrimidine bases are Cytosine and Thymine. These are written as A, G, C, T respectively.

Primary Structure

Because only the nucleobases are different from monomer to monomer, it is sufficient to write the sequence of nucleobases, e.g. ACGGTA, to define the primary structure of a DNA molecule.

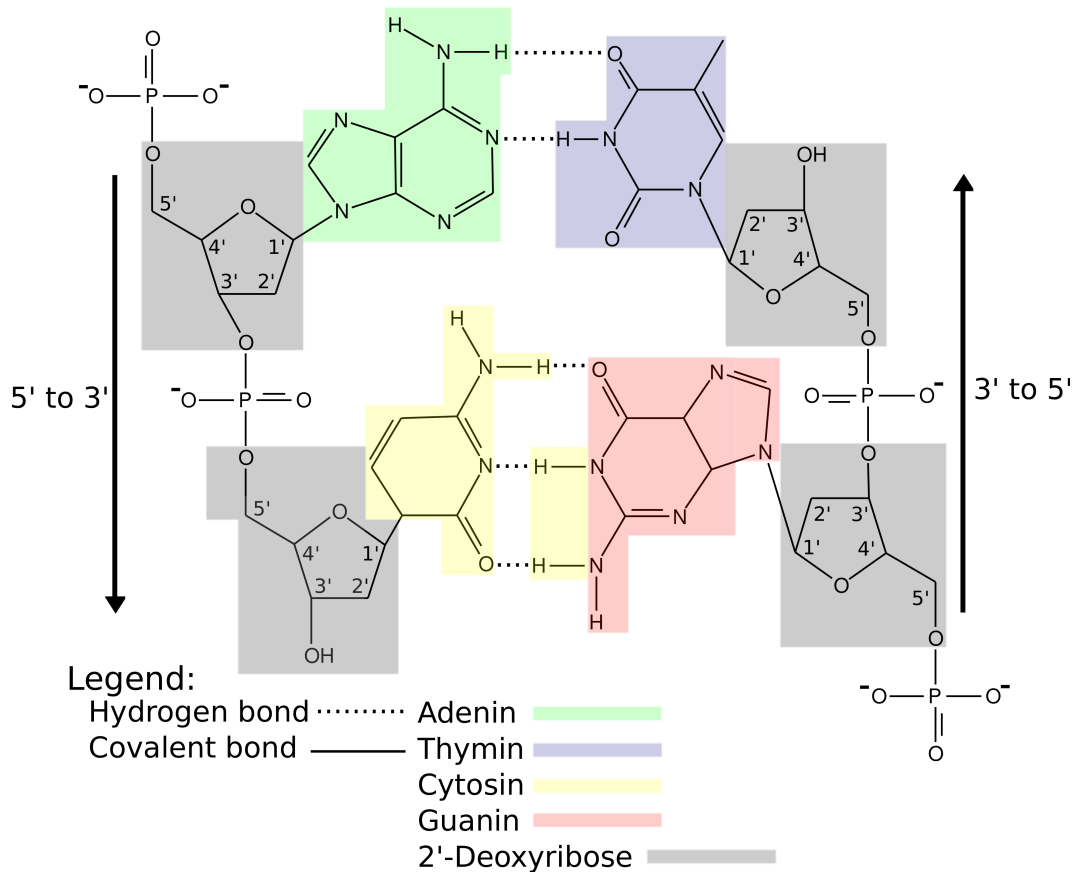


Figure 2.2: DNA double strand showing Adenin forming a base pair with Thymin and Cytosin with Guanin. Directionality of the strands is denoted with arrows. Numbers label the atoms of 2'-Deoxyribose.

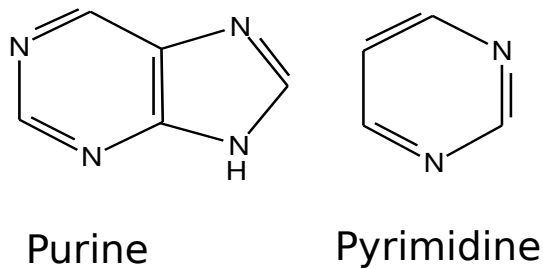


Figure 2.3: Heterocyclic aromatic organic compounds. Purine consists of a Imidazole Ring fused to a Pyrimidine ring.

Secondary Structure

The secondary structure of nucleic acids is defined by intra- or intermolecular base pairing interactions. In biological context cellular DNA is usually in double stranded (Figure 2.2) form, where two DNA polymers are bound to each other via hydrogen bonds between the nucleobases.

The process of forming the double strand is known as hybridization, the re-

versal as melting. The canonical configurations are Watson-Crick base pairs, where Adenine forms base pairs with Thymine (2 hydrogen bonds) and Guanine with Cytosine (3 hydrogen bonds) [10].

Base pairing behavior and the directionality are of importance for the interpretation of the sequence information during transcription and translation.

Figure 2.2 shows the four bases on two different strands forming Watson-Crick base pairs. The directionality of both strands is inverted, the 5' end of one paired with the 3' end of the other one, also known as anti-parallel.

Tertiary Structure

The tertiary structure is defined by the atom-coordinates of the DNA molecule in 3 dimensional space. The DNA strands are wrapped around each other in a helical configuration [11], where the bases are enclosed on the inside of the helix, while the negatively charged phosphate backbone lies on the outside.

The formed structure possesses regularly repeating features like, minor and major groove. These grooves provide surface for interaction with other molecules.

Quaternary Structure

The quaternary structure encompasses the molecules that interact with the DNA molecule, like proteins (e.g. RNA-polymerase, transcription factors, DNA-polymerase), RNA (during transcription) and also DNA (replication and recombination).

In eukaryotic cells the DNA double strands are bound to histone-proteins, which allow a much more spatially condensed configuration of the DNA, known as chromatin. Chromatin controls which parts of the DNA molecule are accessible for transcription [12].

2.1.2 RNA

Primary Structure

Similarly to DNA, RNA is a polymer made of nucleotide monomers (nucleic acid). The monomers also consist of phosphate, sugar (D-ribose) and a nucleobase. Three of the nucleobases are similar to DNA but Thymine is replaced by Uracil (Pyrimidine-base). A further difference is that in contrast to DNA, the RNA molecule is not deoxygenated at the 2' Carbon-atom, as shown in Figure 2.4.

The presence of this hydroxy-group has important implications for the chemical versatility of RNA, which can also function in an enzymatic [13, 14] way often referred to as ribozyme [15].

The primary structure of RNA is the sequence of nucleotides. It can be represented by uppercase letters for each nucleobase (the variable part of the

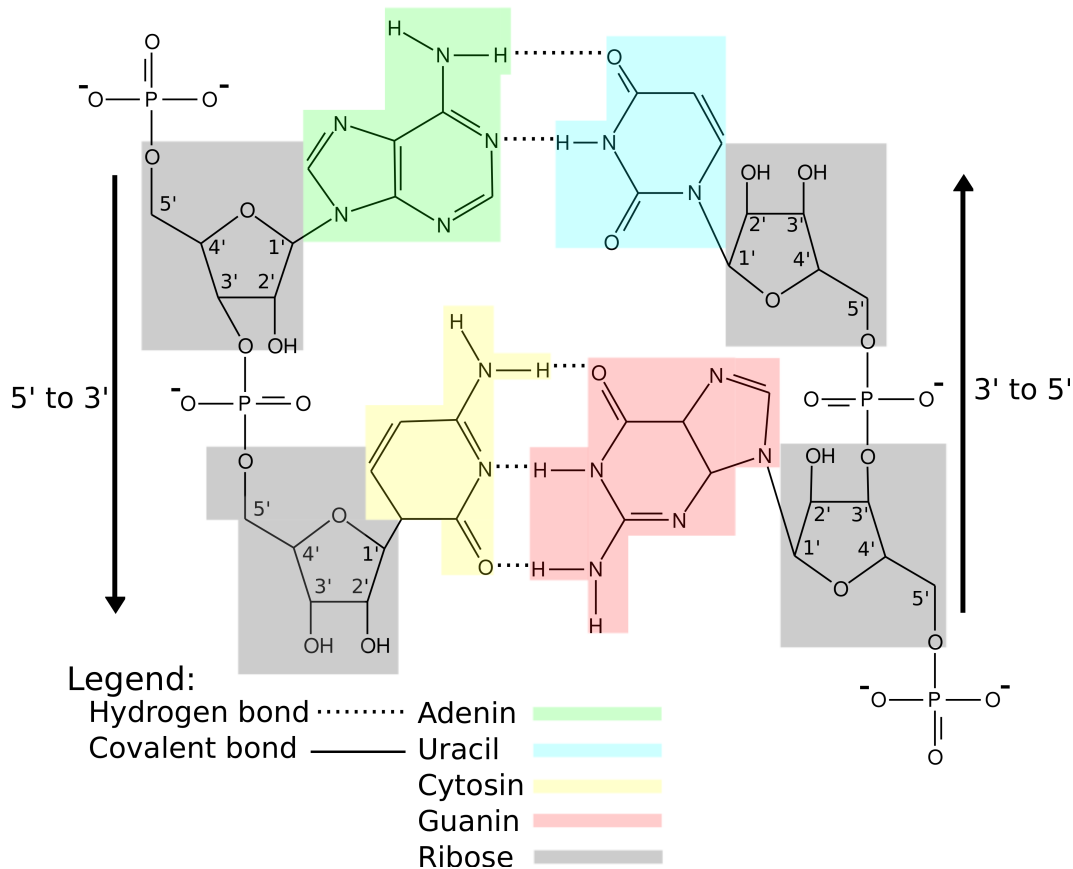


Figure 2.4: RNA double strand showing Adenine forming a base pair with Uracile and Cytosine with Guanine. Directionality of the strands is denoted with arrows. Numbers label the atoms of ribose.

monomers).

Secondary Structure

RNA is, in contrast to DNA, commonly found in single stranded form. This allows the RNA molecule to form intramolecular nucleotide base pairs, yielding complex secondary structures.

While only canonical base pairs were mentioned in context of DNA, also non-canonical base pairs exist and have an important role in tertiary structure formation.

A nucleotide has three edges [16] for interaction as shown in Figure 2.5.

Currently canonical base pairs established over the Watson-Crick edge and the non-canonical Guanine-Uracil Wobble base pair [17] (Figure 2.6) are considered for secondary structure and interaction predictions. The remaining non-canonical base pairs are not considered in the following bioinformatic part, but may be essential for future approaches to RNA-RNA and RNA-Protein

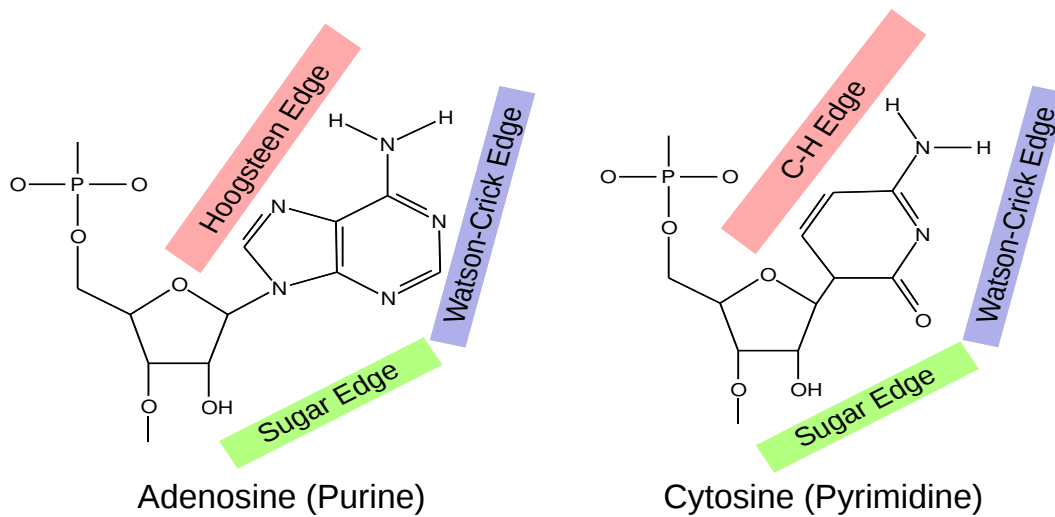


Figure 2.5: Nucleotide interaction can be established between every combination of edges shown above. Edges as recommended by Westhof and Leontis [16].

interactions.

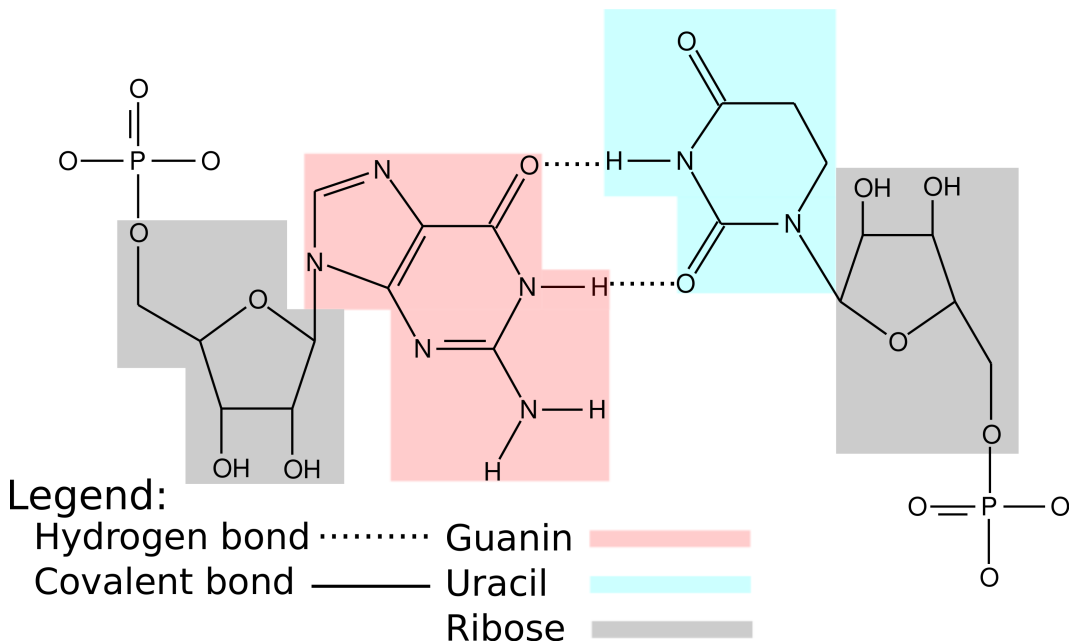


Figure 2.6: Guanine-Uracil wobble base pair is considered non-canonical and establishes 2 hydrogen bonds between the nucleobases.

These base pair interactions allow us to characterize the major motifs of secondary structure [18], see Figure 2.7:

- **Double Helices:**
Double helices are formed by self-complementary base paired regions and are further stabilized by stacking of adjacent base pairs [19]. RNA helices

are in A-conformation, due to their additional 2'-OH group, which results in a deep major and a shallow minor groove [20].

- **Internal Loops:**
Unpaired nucleotides belonging to one (designated bulge) or two strands between duplex regions. If the same number of unpaired nucleotides are present on both strands the internal loop is classified as symmetric, in the other case as asymmetrical.
- **Hairpin Loops:**
The 5' and 3' ends of a double helices connected with a loop of unpaired or miss-matched nucleotides.
- **Junction Loops/Multiloops:**
Three or more double helices with linker sequences of size zero or larger can form intersections designated junction loops or multiloops.

These motifs are often flanked by single stranded so called dangling ends at the 5' and/or 3' end of the sequence.

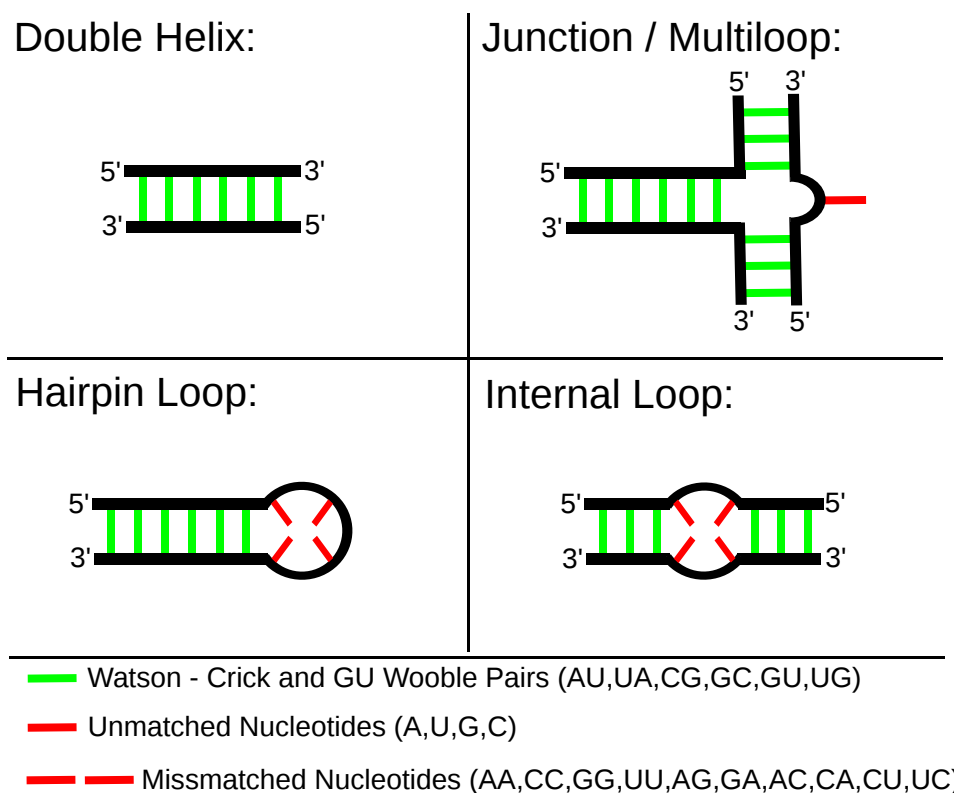


Figure 2.7: RNA secondary structures adopted from [21] and [22]. The representation is inspired by planar graphs (see 3.2.6).

RNA secondary structures can be interpreted as graphs which will be further discussed in the bioinformatic background section [21].

Tertiary Structure

Tertiary structure is defined by the atom-coordinates of the RNA molecule in 3D space. The hierarchical view of RNA folding is a useful abstraction [23] for the bioinformatic approach because it postulates that secondary structures are formed rapidly by stronger non-covalent interactions compared to tertiary structure, which follows slowly and does not induce further changes in secondary structure. This allows to analyze these two abstraction levels independently [24].

Determination of the tertiary structure occurs by:

- Long Range Intramolecular Interactions:
pseudo-knots, ribose zippers, kissing hairpin loops, tetraloop-tetraloop receptor interactions, coaxial helices
- Intermolecular Interactions:
with ligands including metals, small molecules and macromolecules (Protein, DNA, RNA) [18].

Quaternary Structure

The quaternary structure defines the interaction partners of RNA molecules, which are generally other RNAs, Proteins, DNA and small molecules. The most prominent quaternary structure of an RNA can be found in the ribosome [25] which consists of several protein and RNA subunits.

2.1.3 Protein

Proteins, also called polypeptides, are not nucleic acids like DNA or RNA but based on amino acid monomers. The cell deploys proteins to catalyze nearly all reactions of metabolic pathways but also in mediating signals. Proteins are chemically even more versatile than RNA because of the different residues of its monomers. They are relevant for this work because their expression can be regulated by small non-coding RNAs.

Primary structure

The primary structure of proteins is the sequence of its amino-acids, Figure 2.8.

Each amino-acid has an amino- and a carboxy-group between which the peptidic bond is formed, leaving the amino group of the leading amino-acid (N-terminus) and the carboxy-group of the terminal amino-acid (C-terminus) unbound. The protein is synthesized starting at the N-terminus, which has lead to the convention to write amino-acid sequences in the same fashion using either the one or the three letter code shown in Figure 2.8.

Due to the binding of the amino group to the next carbon atom (alpha-carbon) following the carboxy-group the amino-acids found in cells (proteinogenic) are designated alpha-amino acids. The amino-acids are linked covalently, in a

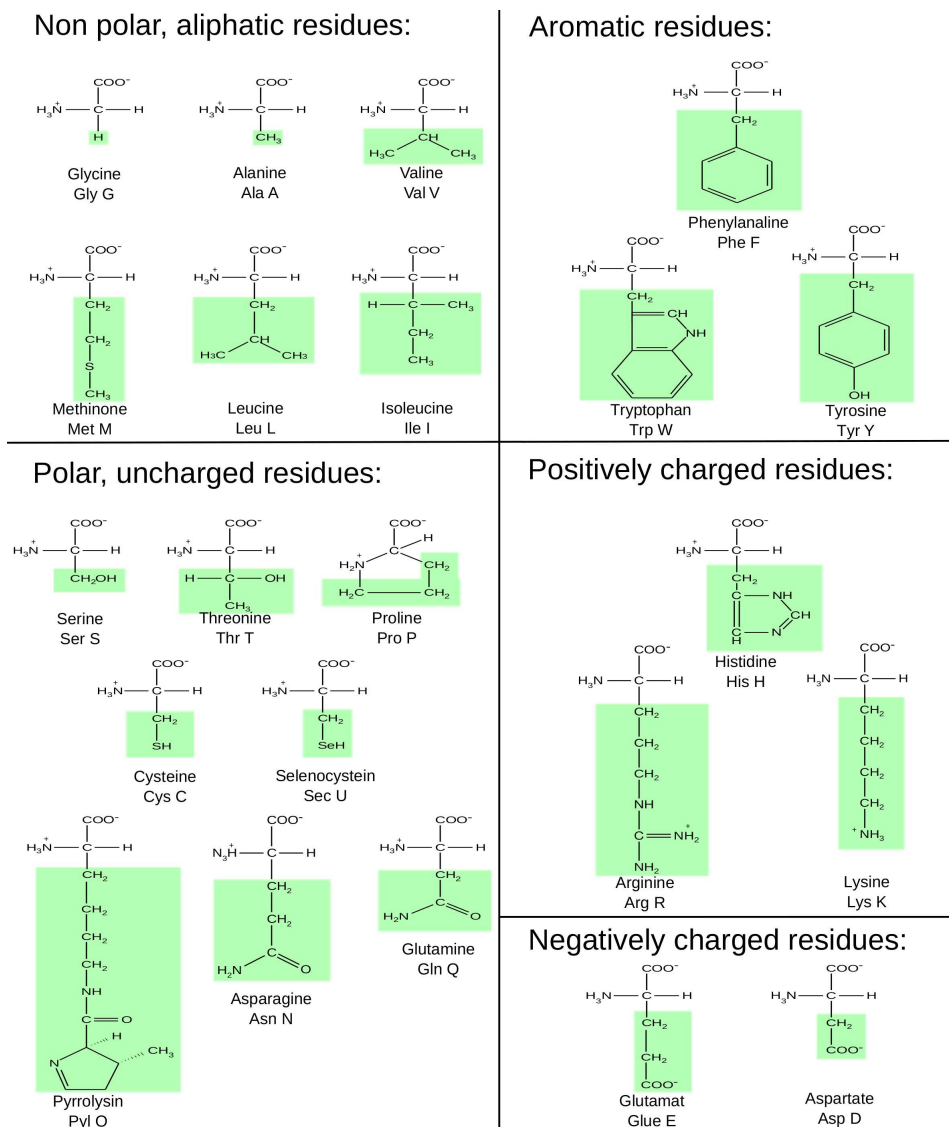


Figure 2.8: Amino acid monomers grouped by chemical properties. The variable side-chains are highlighted in green. Adopted from [26].

linear unbranched manner, over the so called peptide bond, which allows to assign a direction to their sequence (N- to C-terminus).

Figure 2.9 shows a tripeptide, i.e. a protein consisting of three amino-acids.

Secondary Structure

Secondary structures of proteins are established by non-covalent hydrogen-bond interactions between amino and carboxy-groups of the backbone [27]. The most common motifs are the alpha-helix and the beta-sheet. They can be combined among themselves or with others so called super-secondary structures [28], like the coiled-coil.

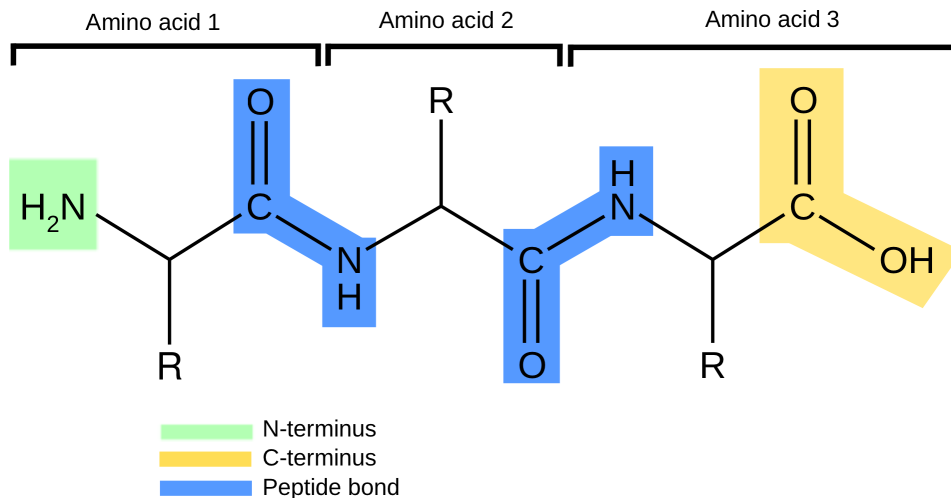


Figure 2.9: Tripeptide consisting of three amino acids connected by two peptide bonds.

Tertiary Structure

Tertiary structure is defined by the atom-coordinates of the protein molecule in 3D space. The reorientation of hydrophobic residues towards the core of the protein is commonly an important driving force [29] in the folding of the protein.

Quaternary Structure

The quaternary structure [30] defines the interaction partners of proteins which are often assembled into larger complexes with other proteins, for example the DNA-polymerase or membrane transporters. Also interactions with DNA and RNA as mentioned before are common.

2.1.4 Transcription

Transcription is the production of RNA by polymerizing nucleotide monomers according to a DNA-template. Both non-coding RNAs and coding RNAs, are assembled by this process. This subsection describes transcription in bacteria, with special focus on *E. coli*. Other species diverge from these mechanisms.

The optimal regulation of transcription is a crucial point in the survival and proliferation of cells [31]. Apart from housekeeping genes which code for enzymes that are always needed to sustain the metabolism the majority of genes is only required in certain conditions, like stress (starvation, temperature). Transcribing currently unnecessary genes is not only a waste of energy and resources for the cell, moreover it could have an effect adverse to the momentary situation.

The transcription unit contains the promoter which is essential for the initiation of transcription, and several coding sequences, each coding for a protein. In the case of several protein coding genes being transcribed to one RNA molecule the transcription unit is called polycistronic, in the case of only a single gene being encoded monocistronic is used. Polycistronic transcription units are nearly exclusively found in bacteria.

Figure 2.10 shows a sequence of DNA also called transcription unit containing all information and signals necessary to be transcribed.

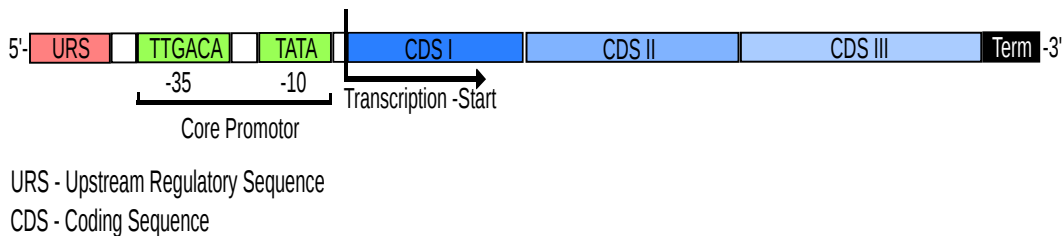


Figure 2.10: Schematic polycistronic transcription unit showing promoter, coding and terminator sequences. This stretch of DNA is transcribed into 1 RNA molecule, but can then be translated to 1 or several distinct protein molecules (polycistronic transcript).

Regions on the DNA strand in 5' direction of the transcription start are known as upstream and in 3' direction as downstream. In contrast to eukaryotic cells bacteria have only one RNA-polymerase. It consists of several protein subunits and transcribes all cellular RNAs. RNA-dependent RNA-polymerases, which have been introduced by bacteriophages, can be found in bacteria [32].

Initiation

Initiation is achieved by binding of the RNA-polymerase to the promoter. There are different kinds of promoters for different types of transcription units. Usually sets of transcription units with the same biological function share the same promoter. The σ -subunit (σ -factor) of the RNA-polymerase that is required for promoter-specificity is variable. By using different σ units depending on the current situation of the cell it is possible to switch whole sets of transcription units and corresponding genes on and off. This situation dependency is also a feature of small RNA mediated regulation.

Three examples from *E. coli* describing a general and two specific σ factors:

- σ^{70} -factor
General factor that binds to promoters of housekeeping genes consisting of the -10 and -35 boxes upstream of the transcription start as shown in Figure 2.10.
- σ^{32} -factor
Specific factor for heat shock proteins

- σ^{28} -factor
Specific factor for motility and chemotaxis

The assembled RNA-polymerase subunits excluding the σ -subunit are designated as core enzyme and including the σ -subunit as holoenzyme.

Besides the core promoter, upstream regulatory sequences can influence transcription. Proteins bound to these sites can down-regulate transcription by e.g. blocking the core promoter, or up-regulate by actively recruiting the RNA-polymerase to the promoter.

By binding of the RNA-polymerase to the promoter, the closed complex is formed and converted to the open complex by melting of the DNA double-strand. This forms the transcription bubble which allows to match complementary nucleotides to the template DNA strand.

Then the first nucleotides can be polymerized which is the onset of elongation. Until the RNA-polymerase has left the promoter region this is referred to as promoter clearance.

Elongation

A fundamental concept is that nucleic acids are elongated by RNA-polymerase from 5' to 3' direction while the template strand is read from 3' to 5' (Figure 2.11). The reading occurs by matching of complementary nucleotides while the elongation is based on covalently linking the matching new nucleotide-triphosphate in an exogenic reaction to the 3'-OH group of the last nucleotide's ribose with elimination of a diphosphate.

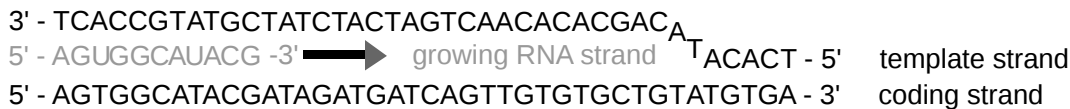


Figure 2.11: Transcription represented by growing RNA that is complementary to the DNA template.

This means that the RNA strand is similar to the coding strand of the DNA (with the exchange of Thymine to Uracil and 2'-deoxyribose being replaced by ribose) but grows by matching it to the template strand. Therefore it is not only necessary to know the genomic coordinates on which a gene or transcript of interest is encoded, but also which strand it is on.

The nucleotides in the growing RNA strand are not always complementary. This error can be corrected by proofreading:

- Hydrolytic Editing [33]:
RNAPolymerase moves some nucleotides back from the end of the growing RNA and removes the miss-matching nucleotide.

- Pyrophosphorolytic Editing [34]:
corrects the last incorrectly added nucleotide by replacement with pyrophosphate.

An other effect resulting from transcription, known as supercoiling, should be mentioned. While the major energy cost of polymerization is contributed by the added nucleotide-triphosphates, additional energy is needed to compensate for compression of the coiling in front of the active RNA-polymerase by releasing tension with an enzyme class called topoisomerase [35].

Termination

Once the termination signal at the end of the transcript is reached there are two different ways to terminate the transcription:

- Rho dependent termination [36]:
The transcribed RNA has binding-sites for the Rho protein which actively pulls the RNA out of the RNA-polymerase. This is achieved by moving along the polymerizing RNA with a speed higher than the growing speed of the chain.
- Rho independent termination [37]:
A stem-loop structure followed by poly-uracil is transcribed. The stem-loop is more stable than the poly-uracil still bound to the DNA which pulls the remaining RNA out of the RNA-polymerase and thereby terminates the transcription.

The products of transcription(mRNA, rRNA, tRNA and small RNA) are discussed in context with translation in the next section.

Degradation of RNAs

The importance of having the necessary genes for the current situation expressed has been mentioned before. But the environment of a cell keeps changing and it is necessary to degrade or disable existing transcripts that are no longer needed. One way to do this is the Degradosome found in *Escherichia coli*, which is a complex formed by three enzymes [38]:

- RNAase E:
Large multidomain protein with ribonucleolytic activity at the N-terminus. The C-terminus binds PNPase and enolase. Serves as scaffold for the complex.
- PNPase:
Polynucleotide phosphorylase with 3' exoribonuclease activity.
- RNA helicase:
DEAD-box RNA helicase which is strongly activated upon binding to RNAase E. Used to unwind structured RNAs.

The half-live of an RNA in *E. coli* is between 30s and 20 min [38].

2.1.5 Translation

Translation is the production of protein by polymerizing amino-acids according to an RNA-template. This subsection describes translation in bacteria, with special focus on *E. coli*. Other species diverge from these mechanisms. Bacterial sRNAs usually influence translation, therefore this step will be described in greater detail. Transcription in bacteria is directly coupled to translation. There are four major components involved in translation:

- **The Ribosome:**
The fully assembled prokaryotic ribosome is designated 70S due to its sedimentation-coefficient of 70 Svedberg units during centrifugation. It consists of a large 50S (containing 23S rRNA, 5S rRNA and about 20 proteins), containing the DC (Decoding center) and a small 30S (containing 16S rRNA and more than 30 proteins) subunit, containing the peptidyl transferase center (PTC) [39]. The catalytic core forming the peptide bond is built from RNAs which classifies the ribosome as ribozyme [40].
- **tRNAs [41]:**
tRNAs short for transport RNAs are RNAs loaded with amino-acids. They function both as adapter by translating the genetic code to amino-acids and as transporter. The different tRNAs share a specific 3 dimensional structure which allows them to fit in the acceptor (A) and peptidyl-site (P) of the ribosome.
- **mRNA:**
The mRNA is transcribed by RNAPolymerase as described above. It serves as template for the translation process. A stretch of mRNA sequence can either be coding or non-coding which is also referred to as UTR (untranslated region).
- **Protein:**
Protein is the product of the translation process.

Before the translation initiation an activation step for the tRNAs is needed, which loads them with the appropriate amino acid.

Initiation

There are three different messenger RNAs shown in Figure 2.12, each representative for a different mode of translation initiation in bacteria [42].

To explain translation the concept of the genetic code needs to be introduced.

Genetic code Translation is based on the genetic code [44], which is used to convert a RNA sequence to an amino-acid sequence. three consecutive nucleotides in the RNA sequence form a so called codon. There are four different nucleotides for each position in the codon sequence resulting in 64 different possible combinations. The genetic code is redundant, as several

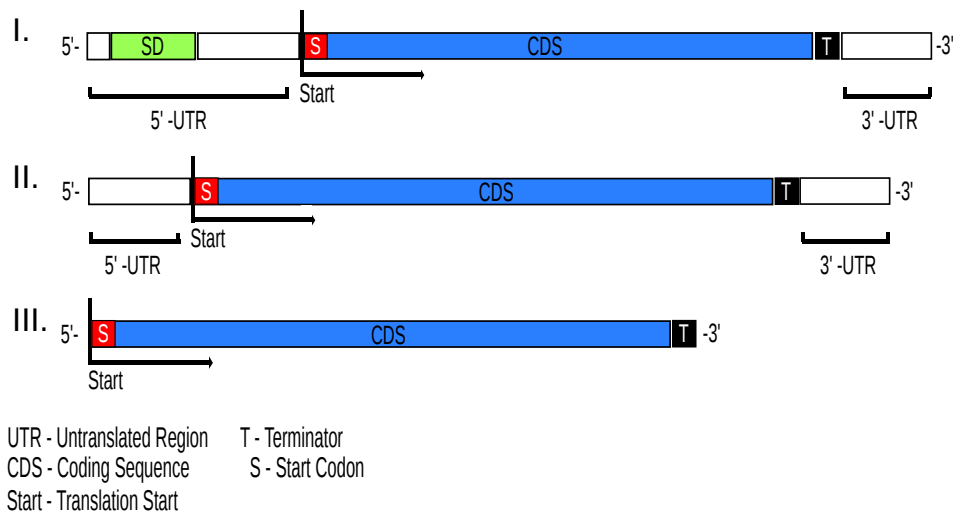


Figure 2.12: Schematic representation of bacterial mRNAs. The mRNA labeled I. features a 5' UTR with a so called Shine-Dalgarno (SD) sequence [43] and is used in the SD-dependent initiation. The transcript labeled II has a short 5'UTR that is utilized in SD-independent Initiation. The III mRNA has no 5'-UTR and is used in leaderless initiation.

codons are interpreted as the same amino acid. Despite this, the code is not ambiguous, as one codon is always interpreted in the same way. Additionally there are codons with special meaning such as start and stop codon. The translation code below (Figure 2.13) is specific for *E. coli*, in other species it is resemblent but not necessarily equal [45].

Two of the amino-acids (Pyrolysine, Selenocystein) mentioned during the description of proteins are not included in this table. They are encoded under special circumstances by STOP-codons [46].

The term reading frame defines which triplets of nucleotides in a specific sequence are interpreted (see Figure 2.14) as codons. Essential for the formation of the initiation complex is the binding of initiator-tRNA (tRNA_i, usually fMet-tRNA) to the start codon.

Shine Dalgarno sequence dependent initiation This is the canonical and most frequently used way of translation initiation (Figure 2.15). First the preinitiation complex is formed which transists to the initiation complex by a process called mRNA adaption [39].

The preinitiation complex is formed from the 30S subunit of the ribosome which is dissociated from the 50S subunit by Initiation Factor (IF) 3. Then IF2 and IF1 subsequently bind to the small subunit. In the next step the initiation-tRNA and the mRNA are assembled into the complex.

The interaction of the mRNA and the 30S subunit is mediated by the formation of an RNA-RNA duplex between the SD sequence and a complementary anti-SD sequence located in the 3' region of the 16S rRNA [43]. The part of mRNA participating in the initiation step is designated translation initiation region

Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met (Start)	ACG	Thr	AAC	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Figure 2.13: Amino acid code adopted from [45] Each of the codons is interpreted by the translation machinery as amino acid or special signal, as shown in the corresponding second column.

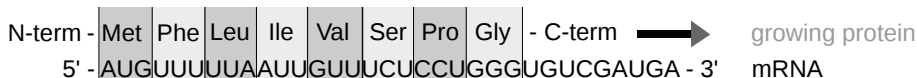


Figure 2.14: Translation represented by matching amino acids of a growing protein with the according codons.

(TIR - Figure 2.16). Several factors marked in Figure 2.16 with upper-case letters influence the strength of a TIR [39]:

- Sequence of the start codon (A)
- Spacing between SD and start codon (B)
- Nonrandom distribution of nucleotides surrounding the TIR (enhancer, C)
- mRNA secondary structure elements present in the TIR (D)

Variation of these parameters can make the binding between TIR and ribosome weaker. In this case, additional interactions, for example between an optional pyrimidine-rich region in the 5'UTR and ribosomal protein S1 [47], are necessary to establish the preinitiation complex.

Once this first preinitiation step is made the mRNA is accommodated in the preinitiation complex by a much slower process. This places the mRNA into the final mRNA channel [39]. Then the start codon-anticodon interaction with the initiator tRNA is formed and the 30S initiation complex is completed. Finally the 50S subunit docks to the 30S initiation complex and IF1 - 3 are

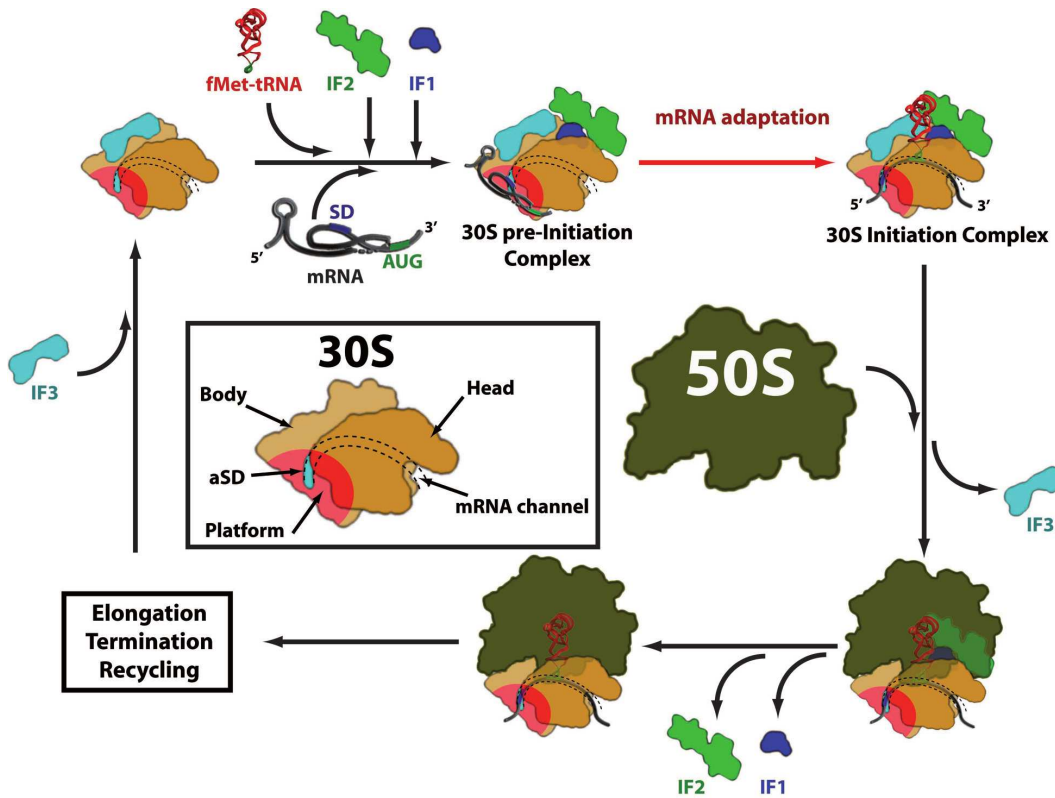


Figure 2.15: The figure shows an overview for translation initiation . The stepwise assembly of the different components is shown. With kind permission from Springer Science+Business Media: [39], Figure 1

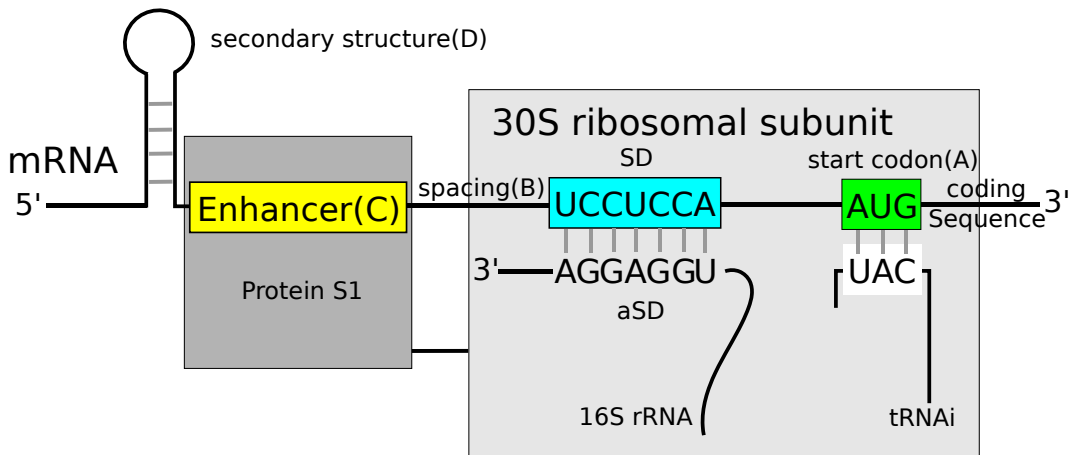


Figure 2.16: Translation initiation region with bound 30S subunit and fMet-tRNA. Secondary structure, enhancer and SD are optional regions. Adopted from [47]

released. The complex is now called 70S initiation complex. The 70S initiation complex is committed to the start codon and tRNA because during the release of the IF2 intrinsic GTP is hydrolyzed to GDP and inorganic Phosphate.

The whole initiation complex assembly is controlled by the three initiation factors:

- Initiation Factor 1 (IF1):
Blocks the acceptor site of the ribosome during initiation [48]
- Initiation Factor 2 (IF2):
Selects for tRNAⁱ binding [49]
- Initiation Factor 3 (IF3):
Discriminates against elongator tRNAs binding instead of the fMet-tRNA^{fMet} [50] Dissociation of S50 subunit.

With the 70S initiation complex established the ribosome can proceed to elongation.

Shine Dalgarno sequence independent initiation A significant part of analyzed prokaryotic genomes has open reading frames with SD free 5' UTRs [51]. There are different mechanisms [42] involved in triggering SD independent initiation, some of them dependent on ribosomal protein S1 [52] or fully assembled 70S ribosomes [53].

Leaderless initiating The assembled 70S ribosome binds the start codon directly. [42] Additional factors can be necessary for initialization. In general leaderless mRNA translation can be initialized in all cells [54], which would also present a means for horizontal gene transfer, mediated by viruses and transposons [55].

Elongation

Figure 2.17 shows a complete cycle of translation elongation [56]. The ribosome has 3 sites in which tRNAs can be bound. In the first round of elongation the initiator-tRNA is bound at the peptidyl (P)-site to the start codon of the mRNA. The acceptor (A)-site that has been blocked by IF2 is now free for docking of new aminoacylated tRNAs. These tRNAs form a ternary complex with Elongation Factor Tu and GTP. The binding of a complementary ternary complex to the A-site triggers GTP-hydrolysis by EF-Tu. This is followed by dissociation of EF-TU + GDP from the complex and accommodation of tRNA. The combination of steric recognition [57] of the tRNA and kinetic proofreading [58] is responsible for low error rates. 2 red arrows indicate possible points for rejection of non-matching tRNAs. Accommodation describes the orientation of the aminoacylated tRNA towards the peptidyl transferase site of the large subunit. Then the amino-acid is cleaved from the initiator tRNA by deacylation and transferred to the tRNA bound in the A-site. Subsequently EF-G + GTP binds and translocates the deacylated tRNA to the exit-site and the peptidyl-tRNA from the A-site to the P-site by hydrolyzing GTP. EF-G dissociates and the ribosome is ready for the next round of elongation, where

the deacylated tRNA bound in the E-site will be ejected and the di-peptide now attached to the P-site tRNA can be further extended [56].

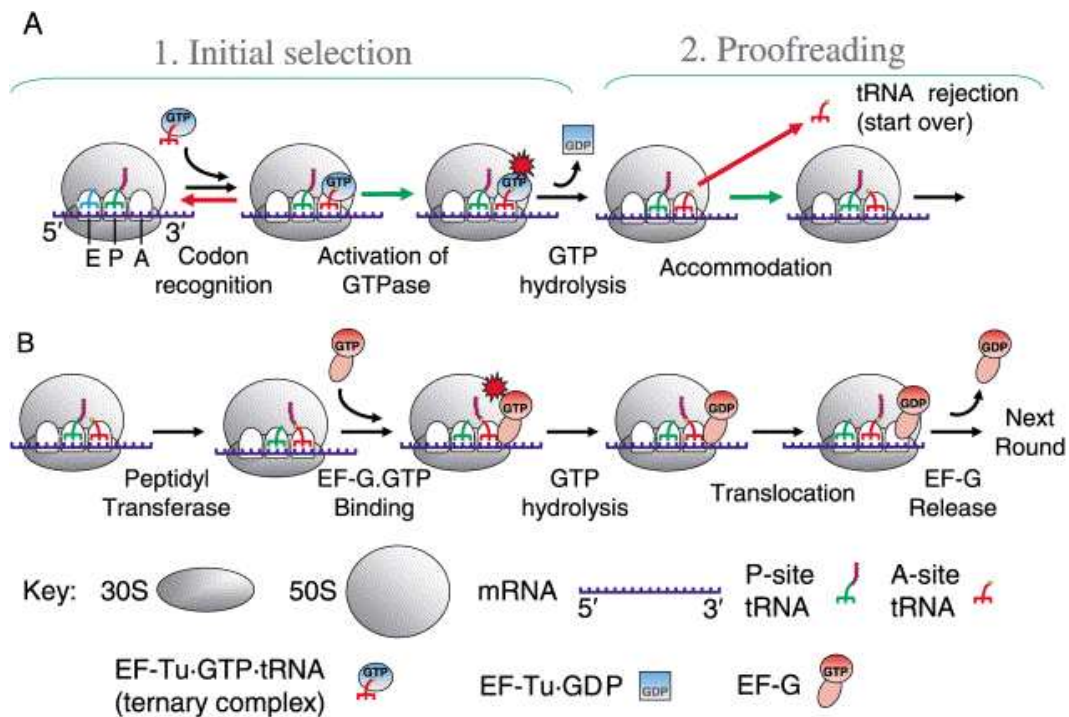


Figure 2.17: Figure shows an overview for translation elongation. A shows the selection of the correct tRNA and proofreading. B shows the peptide elongation by amino acid transfer and the preparation for the next cycle. Reprinted from [56], with permission from Elsevier.

Elongation factors:

- Elongation factor Tu (EF-Tu):
Facilitates binding of correct tRNAs. [56]
- Elongation factor Ts (EF-Ts):
EF-Ts is a Guanine Exchange factor (GEF) and recycles EF-Tu by exchange of bound GDP against GTP. [59]
- Elongation factor G (EF-G):
Catalyzes the translocation step. [49]

Termination

Elongation proceeds until a stop codon is reached (Figure 2.18). The stop codon is recognized by a release factor (RF1 or RF2) which binds to the ribosome. This causes hydrolytic bond cleavage between the protein-chain and the tRNA bound in the P-site. Then optionally RF3 then joins the complex which requires the hydrolysis step of RF1/2 to bind GTP [60]. This causes RF1/RF2

to dissociate from the ribosome and GTP hydrolysis finally causes the release of RF3 [56]. The ribosome still has the mRNA and the tRNA bound and is disassembled (Ribosome recycling) into the large and small subunit (with tRNA and mRNA) by interaction with ribosome recycling factor (RRF) and EF-G + GTP. An alternative pathway with dissociation of the 70S ribosome from tRNA and mRNA has also been reported [61]. In both cases the GTP bound by EF-G is hydrolyzed. The 30S subunit with tRNA and mRNA is bound by IF3 which causes release of the deacylated tRNA. This prepares the ribosome for the next round of translation. Termination factors [56]:

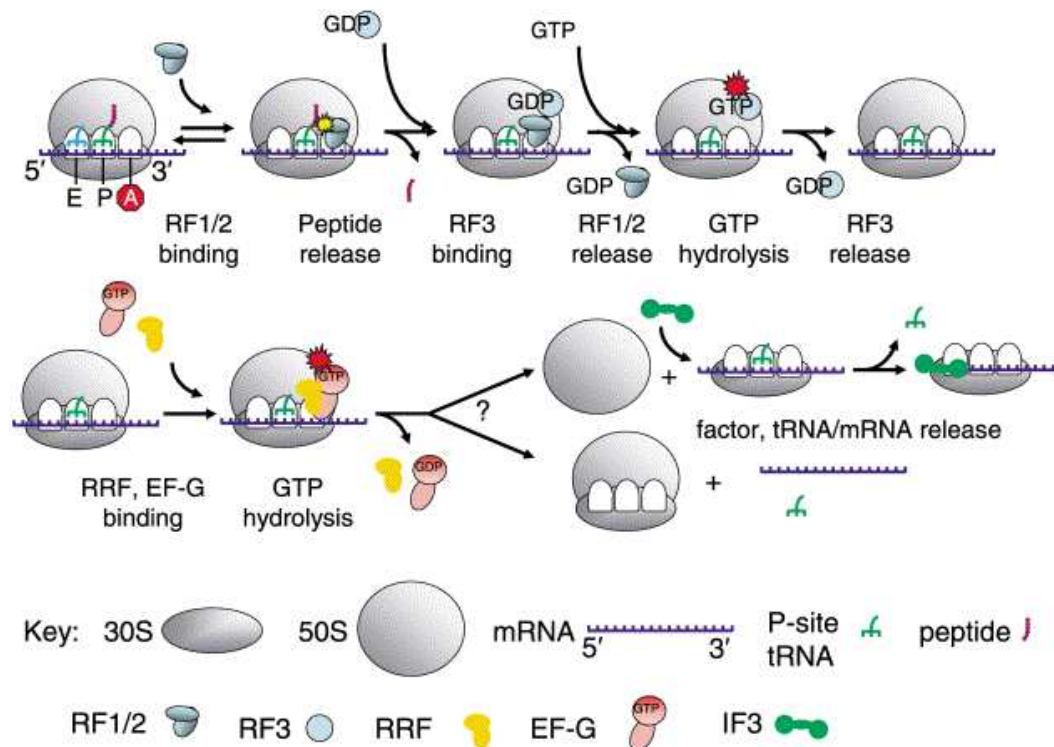


Figure 2.18: Figure shows overview for translation termination. The binding of the release and recycling factors is shown. The ribosome can be disassembled into subunits or stripped of translation factors. Reprinted from [56], with permission from Elsevier.

- Release factor 1 (RF1):
Recognizes UAA and UAG stop codon and belongs to class I release factors.
- Release factor 2 (RF2):
Recognizes UAA and UGA stop codon and also belongs to class I release factors.
- Release factor 3 (RF3):
Is a GTPase and binds GTP when associated to the ribosome upon tRNA deacylation caused by RF1 or 2. GTP uptake leads to RF1/2 release from the ribosome and hydrolysis to dissociation of RF3 itself.

- Ribosome recycling factor (RRF):
Facilitates in combination with EF-G + GTP ribosome subunit disassembly or dissociation of mRNA and deacylated tRNA from S70 ribosomes.

Degradation of proteins

Degradation of proteins that are deteriorated or otherwise detrimental for the cell, is an essential feature. Proteins with incorrect folding or a special tag, for example on the N-terminus (N-end rule) [62], are designated for disassembly. Many prokaryotic and eukaryotic cells rely on proteolytic machines belonging to the AAA+ (ATPase associated with various cellular activities, for example 23S proteasome or Clp protease) super-family [62] for protein degradation, which unfold the target protein in the first step and decompose it in a controlled manner with active groups in their interior. But the effects of protein degradation are more diverse than this, an other illustrative example is the anti-sigma factor. By degrading a transcription factor with a specific protease the transcription of a whole set of genes can be shut down at once. Mechanisms like this can also complement the function of small RNAs as described below.

2.2 Regulatory non-coding RNAs

This work is focused on small non-coding RNAs in bacteria (sRNAs). Non-coding RNAs with comparable function, among others [63], exist in eukarya and will be briefly discussed in the outlook. other ncRNAs, some with comparable function Generally regulatory sRNAs can be divided into two classes depending on their mode of action [64]. tRNAs and rRNAs are excluded. SRNAs binding to Proteins and ncRNAs interacting with messenger RNAs. Both of them enable the cell to react to changing environmental conditions (adaption) by regulating expression of effector and regulatory proteins [64]. Figure 2.19 shows an overview of sRNA action on gene expression [65].

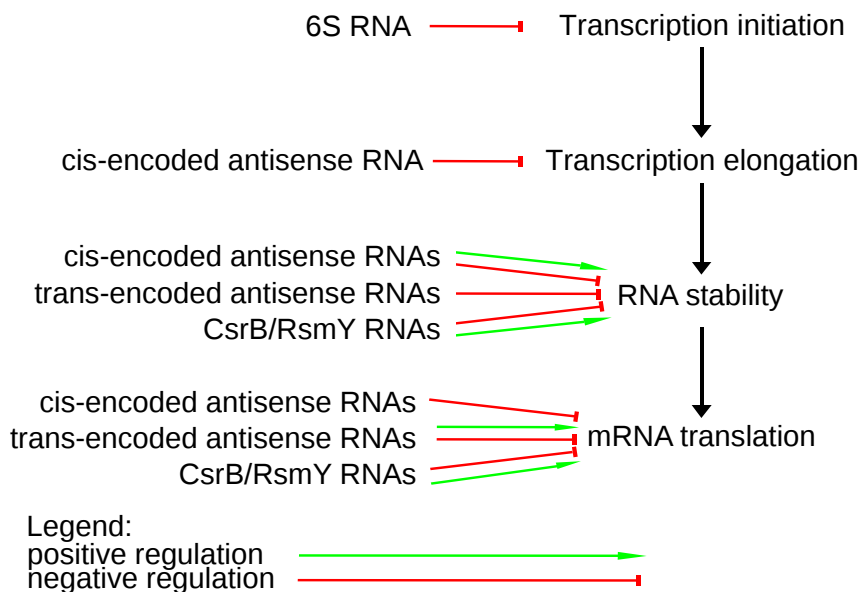


Figure 2.19: Overview of ncRNA action on gene expression adopted from [65].

The focus of this work is on ncRNAs interacting with messenger RNAs.

2.2.1 ncRNAs interacting with messenger RNAs

sRNAs can influence translation and degradation of proteins and can therefore control the total level of concentration of certain proteins. They facilitate a rapid adaption upon transcription because of their relative shortness and because they do not require the additional step of translation to be functional. sRNAs can bind to and affect multiple targets, causing a cascade of signals and therefore ensuring a strong response. Approximately 80 small sRNAs have been identified in *E. coli* and only 60 in other prokaryotic species [7]. There is huge potential for identification and characterization of small RNAs in other bacterial species.

Description

The mRNA binding small regulatory non-coding RNAs of interest for this work, also referred to as sRNAs (small RNAs), have the following features:

- non-coding:
The sRNA transcripts are with few exceptions [66, 67] not coding for proteins. Untranslated regions of mRNAs are, despite their regulatory properties, not considered as sRNAs because they are located on the same transcript.
- small:
Bacterial small RNAs are generally expected to be shorter than 300 nucleotides [68].
- trans-acting:
Regulatory ncRNA that is not transcribed from the same locus as its target mRNA [64].
- cis-acting:
Cis-acting small RNA targets are transcribed from the same locus as their target and trivially predictable through perfect complementarity.

sRNA are either transcribed or processed from already existing transcripts (e.g. GlmZ in *E.coli*) [68, 69].

Mechanism

The sRNA can either have an activating or inhibiting effect on translation. Furthermore the region of base pairing (TIR, coding region, 3'UTR), involvement of Host factor Q (HfQ)-like proteins and potential degradation are of interest. sRNAs bind to mRNAs and form short duplex structures, often including bulges. The shortness of these interactions enables one sRNA to target several cognate mRNAs. Generally translation inhibition/activation is achieved by TIR trapping/freeing [64].

HfQ HfQ is a member of the SM-Protein family [70] and serves as RNA-chaperon. It is highly conserved, forms a hexameric ring structure and promotes the base pairing of several sRNAs to its target, including DsrA, RprA, Spot42 and many more [71]. No HfQ-dependent sRNA-mRNA interactions have been found in Gram positive bacteria. This could be a possible indicator for the presence of a homologous protein or sufficiency of base pairing without a mediator [64].

Translation activation Two classes of translation activation by sRNAs are known [64]:

- Structural changes:

Binding upstream of the TIR on the 5'-UTR causes an increase in structural accessibility of the Shine-Dalgarno sequence, enabling translation (Figure 2.20). An example for this mechanism is the trans-acting sRNA DrsA found in *E. coli* [72].

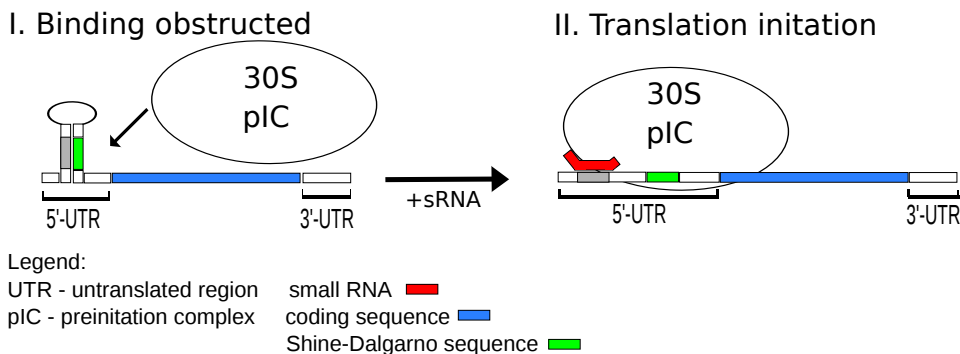


Figure 2.20: Translational activation by increase of structural accessibility.

- Stabilization:

sRNA binding blocks RNase activity (Figure 2.21). In case of *E. coli* GadY base pairing to the 3'UTR of *GadX* mRNA, which is of variable length due to a missing Rho-independent terminator hairpin, degradation of the transcript by RNAseII or polynucleotidephosphorylase could be inhibited by forming a duplex [73].

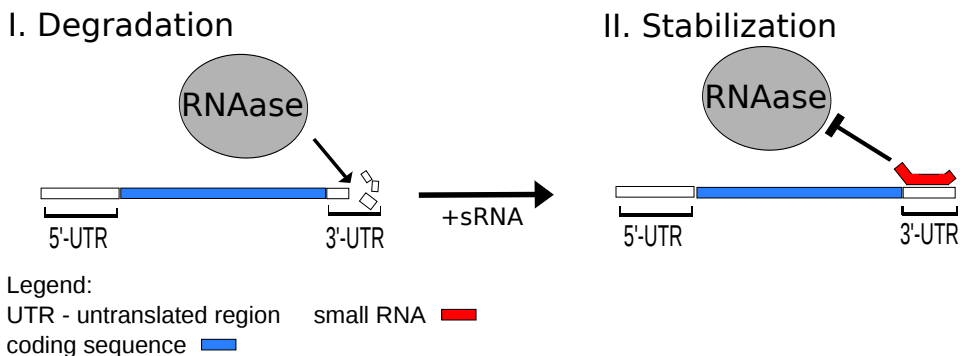


Figure 2.21: Translational activation by stabilization.

Translation inhibition Reported mechanisms for translation inhibition are [64]:

- Base pairing with functional regions of mRNA:

TIR/SD:

Base pairing overlapping with the initiation region blocks binding of the initiation complex (Figure 2.22). OxyS sRNA and *fhlA* in *E. coli* is regulated in this manner [74]. Other examples include pairing with a C/A-rich regions upstream of the TIR (*GcvB* [75]), possibly acting as enhancer.

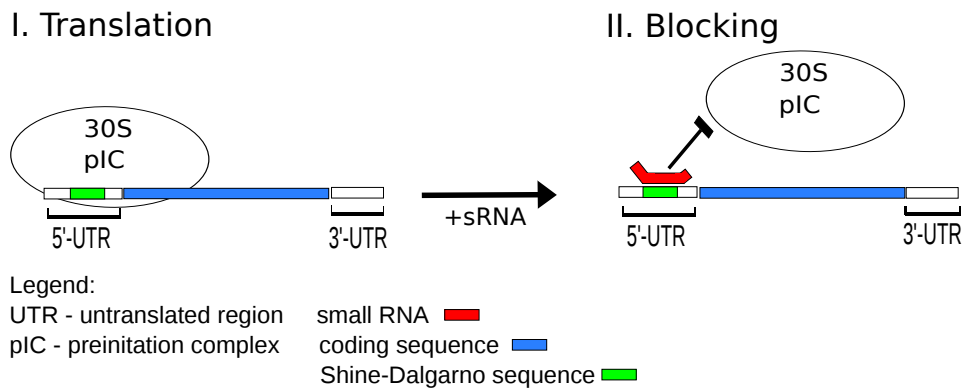


Figure 2.22: Translational inhibition by blocking functional regions.

- Structural changes:
Pairing of SR1 to the coding region of AhrC [76] in *E. coli* causes structural changes in TIR, blocking the binding of the ribosome (Figure 2.23).

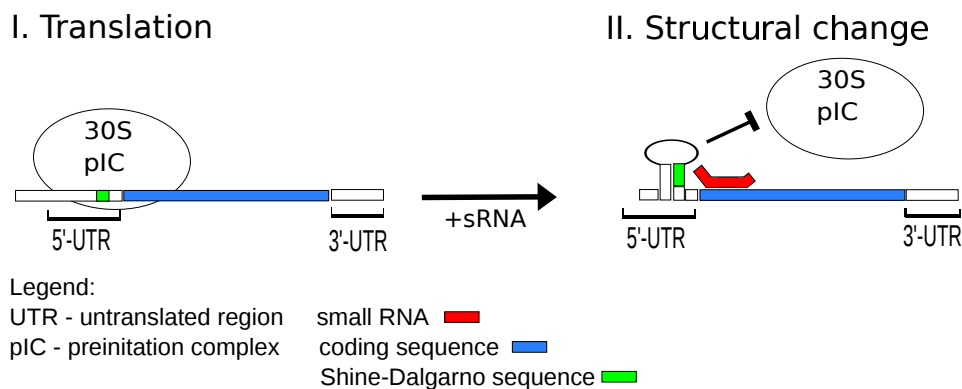


Figure 2.23: Translational inhibition by structural changes.

- Degradation:
The binding of sRNAs is sufficient to inhibit translation, but additional degradation of the RNA, by creating binding motives for RNases, is possible and makes the inhibition irreversible [77] (Figure 2.24). RNase III and RNase E are associated with this mechanism. In case of RNase E a complex with Hfq and ncRNA can be formed for specific degradation [64].

2.2.2 ncRNAs interacting with proteins

ncRNAs can bind to proteins to cause a regulatory effect. Protein/RNA interactions, due to their complexity, are difficult to model and therefore not considered in this work. While ncRNAs interacting with mRNAs influence translation, this can also directly affect transcription. An example from *E. coli* is transcription reprogramming by SsrS (6S RNA) [64]. As described

I. Translation

II. Degradation

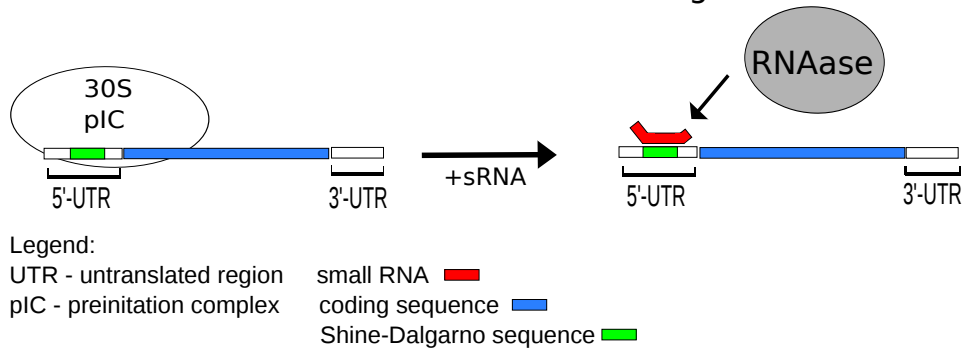


Figure 2.24: Small RNA mediated RNA degradation.

above the RNAPolymerase core enzyme gains promoter specificity by forming a holoenzyme with a σ -factor. SsrS binds to the σ -factor 70-holoenzyme that provides affinity to housekeeping gene promoters and alters the subset of genes transcribed [64]. The SsrS-sensitive promoters are down-regulated during the stationary phase. The conserved secondary structure of SsrS resembles the 'open complex' of translation initiation, which leads to competition between SsrS and promoters for binding with the RNAPolymerase holoenzyme. This similarity also enables the holoenzyme to transcribe a short stretch of SsrS to synthesize product RNA (pRNA), as soon as sufficient nucleotides are available on outgrowth of the stationary phase [78].

3 Bioinformatics

RNApredator uses RNApIex to predict RNA-RNA interactions in silico, based on structural accessibilities calculated by RNApIfold. RNApIex and RNApIfold build on concepts drawn from RNA folding. The focus of this section will be on algorithms for RNA-RNA interaction.

3.1 RNA-Bioinformatics

The high structuredness of RNA secondary structure (compared to, say, proteins) allows for efficient modeling approaches to happen on this level instead of having to model full three-dimensional molecules [24].

- Primary structure:
The sequence of nucleotides can be treated as a sequence of symbols. The information contained in primary structure is sufficient to address a lot of questions, but lacks structural information that is highly relevant for biological function.
- Secondary structure:
Secondary structure can be represented as list of base pairs or as an outer-planar graph assuming that no pseudoknots are present [79]. The vertices of the graph represent the nucleotides, while the edges represent backbone bonds and base pairs.
- Tertiary structure:
Tertiary structure itself can mean a full atomic-scale model, but often reduced representations are used. As the determination of tertiary structure is a very hard problem for RNA, work has focused on secondary structures [24].

3.2 RNA Folding

RNA folding based on minimal free energy provided the foundation for the algorithms used for RNA-RNA interaction prediction used by RNApredator.

Generally RNA folding of secondary structures can be based on comparative sequence analysis or free-energy based [80]. The first algorithms for secondary structure prediction were based on a physics-based model [81, 82]. Another algorithm that relied on maximization of base-pairs using an dynamic programming approach [83] followed. These strategies were combined into RNA

folding that considers thermodynamic stability and stacking energies, while using a dynamic programming approach [84].

The folding recursion [85] and energy parameters presented are as implemented in the *Vienna RNA package* [86].

3.2.1 Additive energy model

The standard additive energy model is used to compute the energy of a given structure. The bounded faces forming the unique minimum cycle basis of the outer-planar graph [85] used to represent the RNA molecule are called loops in this context. Loops form the units of the additive energy model and have a direct biophysical interpretation as entropically destabilizing elements or stabilizing stacked base pairs [85].

The RNA sequence x considered has a length of n . At sequence position k the nucleotide is denoted by x_k . The subsequence (x_k, \dots, x_l) is represented by $x[k,l]$. In the folding recursions below we will require energy parameters for hairpin loops \mathcal{H} , interior loops \mathcal{I} and multiloops \mathcal{M} . These depend on the loop type, length and sequence and have been measured with melting experiments [87, 88]. Visualizations of the loop types are shown in Figure 2.7. The closing pair (k, l) uniquely determines the hairpin loop \mathcal{H} . This base-pair and the number of unpaired nucleotides contribute to the parameter \mathcal{H} .

The interior loop \mathcal{I} is determined by two enclosing base pairs. Its energy parameter consists of the base pair energies and the energy of the unpaired bases enclosed by them.

Multiloops \mathcal{M} have 2 or more branches B not including the branch with the enclosing base pair. \mathcal{M} is constructed from contributions by branches, unpaired bases enclosed by the branches and a energy contribution to close the complete loop.

Despite being implemented in the *Vienna RNA package* [86] dangling end contributions are not included in the recursions below for reasons of clarity.

3.2.2 Folding recursion

RNA Folding is an essential precursor for physics-based RNA-RNA interaction prediction algorithms. The possible set of structures of the RNA molecule is decomposed into a set of substructures that are defined on subsequences. This energy-directed folding is solved by dynamic programming algorithms. The decomposition is conducted in a way that each possible structure is counted only once, which is fundamental for sub-optimal folding and computation of the partition function [89]. For linear molecules the *Vienna RNA package* folding algorithm computes the following arrays for $i < j$:

$x[i..j]$ subject to the constraint that $x[i, j]$ is part of a multiloop and has exactly one component, which has the closing pair i, h for some h satisfying $i \leq h < j$. In summary, the recurrences to compute the minimal free energy folding algorithm [84] for linear RNA molecules as implemented in the *Vienna RNA*

F_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$
C_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$, subject to the constraint that i and j form a base-pair
M_{ij}	free energy of the optimal substructure on the subsequence $x[i..j]$, subject to the constraint that $x[i, j]$ is part of a multiloop and has at least one component, i.e., a sub-sequence that is enclosed by a base-pair
M_{ij}^1	free energy of the optimal substructure on the subsequence

Table 3.1: Arrays used in the folding recursion, taken from [85]

package [86] are:

$$\begin{aligned}
F_{ij} &= \min\{F_{i+1,j}, \min_{i < k \leq j} (C_{ik} + F_{k+1,j})\} \\
C_{i,j} &= \min\{\mathcal{H}(i, j), \min_{i < k < l < j} C_{kl} + \mathcal{I}(i, j; k, l), \min_{i < u < j} M_{i+1,u} + M_{u+1,j-1}^1 + a\} \\
M_{ij} &= \min_{i \leq u < j} (u - i)c + C_{u,j} + b, \min_{i < u < j} M_{i,u} + C_{u+1,j} + b, M_{i,j-1} + c \\
M_{ij}^1 &= \min\{M_{i,j-1}^1 + c, C_{ij} + b\}
\end{aligned}$$

Table 3.2: Folding recursion adopted from [85], each line of the table is showing a part of the structure decomposition

The recursion is initialized as follows $F_{ii} = 0, C_{ii} = M_{ii} = M_{ii}^1 = +\infty$. The memory consumption is $\mathcal{O}(n^2)$. Restriction of total interior loop length to $(j - l - 1) + (k - i - 1) \leq 30$ leads to a time complexity of $\mathcal{O}(n^3)$. Additional restrictions are necessary to consider only canonical structures with this algorithm, as implemented in the *Vienna RNA package* [86]. For details consult [85]. Once the minimum free energies are computed, structures can be obtained with backtracking. The structure is expressed as list of base-pairing edges.

3.2.3 Partition function

The partition function is a necessary prerequisite to compute structural accessibilities, which improve the sensitivity of RNA-RNA interaction predictions. The partition function Z yields information for the ensemble of RNA structures. The frequency of specific secondary structures or the probability of certain bases forming pairs can be computed with it.

In equilibrium, a specific structure S occurs proportional to its Boltzmann factor $\xi = \exp(-E(S)/RT)$ [24]. The partition function (see Equation 3.1) represents the ensemble by summing up the Boltzmann factors of all (secondary) structures for the sequence.

$$Z = \sum_S \exp(E(S)/RT) \quad (3.1)$$

T is the absolute ambient Temperature in Kelvin and R is the molar gas constant. As additivity of free energy causes multiplicativity of the contributions

to the partition function [90] we can transform the folding recursion (see 3.2.2) to calculate the partition function. This is done by replacing maximum operations with sums, sums with multiplications and energies with Boltzmann factors [24]. It is essential that structures are only considered once in the partition function.

3.2.4 Base pairing probability

For RNA-RNA interactions it is of great interest in which conformation the molecules reside before interaction. The partition function allows us to compute the probability for a single specific base-pair (i, j) in the sequence $[i, j]$. The original strategy [91, 90] to compute the probability p_{ij} that i and j pair in thermodynamic equilibrium is as follows [91]:

$$p_{ij} = \frac{Z_{1,i-1} \hat{Z}_{i,j} Z_{j+1,n}}{Z_{1,n}} + \sum_{k < i} \sum_{l > j} p_{kl} \Xi_{ij,kl} \quad (3.2)$$

The partition function Z_{ij} represents all secondary structures on the interval $[i, j]$, while \hat{Z}_{ij} denotes the partition function with the constraint that i and j base-pair. $\Xi_{ij,kl}$ is the ratio of $\hat{Z}_{ij,kj}$ and $\hat{Z}_{k,l}$ with the constraint that both i, j and k, l base-pair [91]:

$$\hat{Z}_{ij,kl} = Z_{k+1,i-1} \hat{Z}_{ij} Z_{j+1,l} \xi_{kl} \quad (3.3)$$

Energies depend only on individual base-pairs in the simplest case, where ξ_{kl} is the Boltzmann factor for the closing base pair (k, l) [91]. Implementations consider the full energy model (see 3.2.2) [87, 89], not the minimal example given above.

Instead of directly using this approach, we used fixed sequence window recursions as implemented in RNAPfold [91].

3.2.5 RNAPfold

The simplest way to compute accessibilities is to forbid base-pairing for a certain stretch of nucleotides and divide the resulting restricted partition function by the unrestricted one [92]. This approach has a time complexity of $\mathcal{O}(n^5)$ for all n^2 possible intervals, where n is the sequence length. Stochastic sampling of structures to calculate accessibilities introduces sample errors, but requires only $\mathcal{O}(n^3)$ [93].

The RNAPfold approach runs in the same efficiency class $\mathcal{O}(n^3)$ but without sampling errors [94].

RNAPfold uses a localized partition function $Z_{ij}^{u,L}$ over all secondary structures on the interval $[i, j]$ for the folding of sequence window $[u, u + L]$. $\hat{Z}_{ij}^{u,L}$ additionally has the constraint that i and j pair. $p_{ij}^{u,L}$ is the probability of base-pairing between i and j , for folding this window. $\hat{Z}_{ij}^{u,L}$ is independent of external

structures if the subsequence is part of the folded sequence window [91]:

$$Z_{ij}^{u,L} = \begin{cases} Z_{ij} & \text{if } [i, j] \subseteq [u, u + L] \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

$p_{il}^{u,L}$ is also 0 if $[i, j] \subseteq [u, u + L]$. This enables the restriction of Equation 3.2 for a sequence window $[u, u + L]$ [91]:

$$p_{ij}^{u,L} = \frac{Z_{1,i-1}^{u,L} \hat{Z}_{i,j}^{u,L} Z_{j+1,n}^{u,L}}{Z_{u,u+L}} + \sum_{k < i} \sum_{l > j} p_{kl}^{u,L} \Xi_{ij,kl}^{u,L} = \frac{Z_{u,i-1} \hat{Z}_{i,j} Z_{j+1,u+L}}{Z_{u,u+L}} + \sum_{k < i} \sum_{l > j} p_{kl}^{u,L} \Xi_{ij,kl}^{u,L} \quad (3.5)$$

This yields a table with $p_{ij}^{u,L}$ that can be further processed.

To use the RNApIex we have to transform the probabilities to energies G_{ij} . This is possible by considering the following equation [95] that gives the energy for an unpaired binding motif, which is substituted by its localized equivalent from RNApIfold $p_{ij}^{u,L}$ [91].

$$\Delta G = (-1/RT)(\ln(Z_u[i, j]) - \ln Z) = (-1/RT) \ln p_{ij}^{u,L} \quad (3.6)$$

We get a table with opening energies for all nucleotides that can be considered in RNA-RNA interaction algorithms.

3.2.6 Visualization of secondary structures

The formulation of secondary structures as different data-structures allows a variety of graphical representations. Loop and stacking regions are of special significance for minimal free energy folding (see 3.2.1).

The representations shown in Figure 3.1 are useful to convey different kinds of information. The secondary structure graph shows a specific structure from the ensemble of possible structures, coloring of the vertices enables visualization of base pairing-probabilities or positional entropy. The mountain plot is useful for comparison of large structures, while the dot plot shows base pairing probabilities over the whole ensemble.

The string in dot-bracket notation has dots for unpaired nucleotides and parentheses for the paired ones. The base-pair is established between the matching parentheses.

3.3 RNA-RNA Interaction

This section will give an overview of algorithms for RNA-RNA interaction and will outline the choices made for the pipeline of RNApredator. Generally one of the nucleotide sequences is designated as query, the other one as target. The query is then tested against multiple targets.

3.3.3 RNA Hybridization

Performing interaction studies with long(genomes) or multiple sequences (transcriptomes) requires a low time complexity. This was achieved by neglecting intramolecular base pairing. RNAhybrid [103] and RNAduplex are examples based on RNA folding algorithms, while bindigo relies on an adaptation of the Smith-Waterman sequence alignment algorithm [104]. Runtimes of $\mathcal{O}(n * m * L^2)$ (folding, with L as max. loop-length) or $\mathcal{O}(n * m)$ (alignment) are possible with this strategies [103].

3.3.4 RNAup

RNAup in contrast considers the energy of intramolecular base-pairing ΔG_u of the larger target molecule and then uses the RNAhybrid approach to compute the hybridization energy ΔG_h [95]:

$$\Delta G = \Delta G_u + \Delta G_h \quad (3.7)$$

RNAup allows the binding to any kind of loop, in contrast to RNAduplex, which only allows the exterior loop of the concatenated sequences, but is limited to one binding site [101]. It has a time complexity of $\mathcal{O}(n * m^5)$ and a memory consumption of $\mathcal{O}(n * m^3)$.

3.3.5 RNAPlex

The methods described before are either fast and imprecise or slow and precise. RNAPlex uses a position depended per-nucleotide penalty to mimic the competition between intra- and intermolecular base pairing. The accessibility profiles from RNAplfold[94] are used to derive the penalties. For genome screening studies it is of advantage to precompute and store the accessibility profiles, because the increase in precision does not increase the time complexity.

RNAPlex is based on RNAduplex but uses a simplification of the energy model, where loop energy is an affine function of the loop size instead of a logarithmic one [101]. This results in a time complexity of $\mathcal{O}(n * w)$ and a memory requirement of $\mathcal{O}(l^2)$, where l represents the maximal hybridization length. Like RNAup, RNAPlex is limited to one binding site.

The newest version of RNAPlex allows the consideration of multiple alignments in the target prediction, which increases specificity[105].

3.4 Gene ontology

The Gene Ontology (GO) project (<http://www.geneontology.org/>) is developing ontologies that provide a systematic language for consistent description of genes and their products [106]. It is part of the Open Biomedical Ontology

(OBO) and shares resources, like tools for annotation (OBO-Edit)[107] and principles with OBO [108].

3.4.1 Ontologies

Proteins or genes are described by the term 'function', which is sometimes used to define biochemical activity, but in other cases biological goals and cellular structure [109]. Gene ontology therefore uses attributes from three independent ontologies (Molecular Function, Biological Process, Cellular Component). The ontologies are networks of unique nodes (GO-term) with defined connections with one or more nodes or more precisely a directed acyclic graph [109]. The GO-terms serve as attributes for genes or gene products and are more or less detailed, depending on the state of knowledge or annotation. The attributes from three ontologies can be assigned independently to a gene or gene product and are common to all organisms.

Molecular function defines the biochemical activity of a gene product. Only the activity or potential activity is described, no information about when and where it occurs is provided.

Biological process defines the biological objective that the gene or its product contribute to. One or more molecular functions are combined in an ordered way to achieve a biological process.

Cellular component is describing the location in the cell the gene or its product is active in.

3.4.2 GO-term

Each GO term has a unique identifier and is connected to its parents, which offer a broader description and its children, which offer a more specific description of either a Molecular Function, a Biological Process or a Cellular Component. Many of such GO-terms can be associated with a gene or its product.

The description of an GO-term contains its unique identifier (accession), the ontology it belongs to, synonyms including exact and related ones and alternative unique identifiers, a definition (e.g. the educts and products of the catalyzed reaction in case of a molecular function) including the source, a comment, the subset the terms belong to and community comments.

4 Methods

RNApredator [1] provides a web service for mRNA and sRNA interaction prediction. The methods described are essential for RNApredator and shall provide an overview for the steps of the pipeline. The first step (Data preparation) requires a lot of computational resources, but can be computed without user input. This preprocessing saves time during target prediction, which is performed with the precomputed mRNA data. The results of target prediction are evaluated and interesting interactions can be further analyzed.

4.1 Data preparation

Hypothetical transcript construction allows to consider not only the coding sequence, but also the untranslated region of the transcript. Structural accessibility is calculated based on hypothetical transcripts and allows to include intramolecular binding of the interaction partners before computing target prediction.

4.1.1 Hypothetical transcript construction

Data concerning transcript sequences is only obtainable for a few selected species. As described in the biological background (see Subsection 2.2.1) interactions with sRNAs often occur in the 5'-UTR of mRNAs. For the majority of species only the coding sequences of the genes are available from databases.

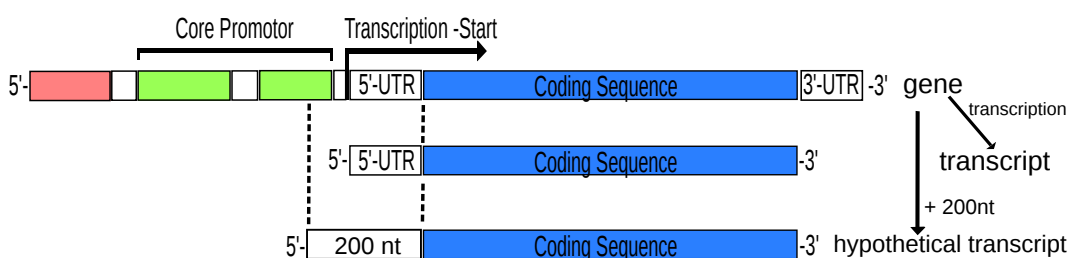


Figure 4.1: Construction of a hypothetical transcript from a genomic sequence (gene) and comparison to a transcript. The interrupted vertical lines denote the 200 upstream nucleotides added to the coding sequence.

As a compromise 200 nucleotides directly upstream of the translation start were concatenated with the coding sequence to serve as hypothetical transcript. Translation initiation depending on the Shine-Dalgarno sequence is the

most common (see Subsection 2.1.5) in bacteria and is covered by this 200nt region. This was done for all genes of every bacterial species available via NCBI (<ftp://ftp.ncbi.nih.gov/genomes/>) for download. In total 1 183 bacterial species with 2 154 genomes/plasmids and 3 759 619 genes were processed.

Developing a model that predicts more accurate transcript sequences would have been too time consuming at this point, but remains a goal for the future.

The genome and coding sequence data is distributed in several different files in fasta format. For the coding sequence the sense strand is provided, which is sequence-identical to the mRNA. In case of the genome only the (plus) strand is provided which is not the sense strand for all genes. To include the 200 upstream nucleotides for these genes the complementary region on the minus strand was computed.

4.1.2 Structural accessibility calculation

RNAplfold [91] (see Subsection 3.2.5) was used to compute the structural accessibilities for all hypothetical transcripts. The adjustable parameter winsize was set to 240, span to 160. Winsize defines the size of the window for which mean probabilities are computed, span allows only pairs (i, j) with $j - i \leq \text{span}$. RNAplfold outputs the logarithm of the mean probabilities that regions of length 1 to 30 (-u 30) are unpaired.

Even with the lower time consumption of the localized approach, this step was consuming approximately 2000h of computation time on Intel Xeon X5550 cores clocked at 2.67 Ghz.

The following Listing 4.1 shows an example structural accessibility output file from RNAplfold.

```
#opening energies
#i\ $ 1=1 2 3 4 5 6 7 8 9 10
1 0.01470122 NA NA NA NA NA NA NA NA
2 0.009218787 0.01681561 NA NA NA NA NA NA
3 0.002687764 0.01189638 0.01948257 NA NA NA NA
4 0.005022743 0.006299687 0.0155313 0.02311421 NA NA
5 0.008950001 0.01180837 0.01306088 0.02235408 0.02992067
6 0.0159508 0.01980088 0.02268747 0.02392594 0.03325052
7 0.01448151 0.01786842 0.02167088 0.02453824 0.02577072
8 0.009211326 0.01946468 0.02279271 0.02646765 0.02935033
9 0.004709575 0.01004038 0.02027955 0.02358692 0.02723467
10 0.001636226 0.006154286 0.01136356 0.02159105 0.02488072
```

Listing 4.1: Structural accessibility output, showing the logarithm of the mean probabilities that a region of length 1 up to 30 is unbound.

The precomputed structural accessibility files use about 683GB of disk space. Compression with gzip lowered the space consumption approximately by 30% but requires either decompression before target prediction, or a target prediction implementation that features reading directly from compressed structural

accessibility files.

The accessibility of the sRNA is computed with RNAup [95] with the length of the unstructured region set to 30 or the length of the sRNA if shorter. RNAup accessibility calculation uses a more complex energy model.

4.2 Target prediction

RNA-RNA interaction prediction provided by RNApredator is used to find the best putative mRNAs targeted by sRNAs. The targets of an mRNA can give a clue about the biological function of an sRNA.

Target prediction uses the sequence data from hypothetical transcript construction (see Subsection 4.1.1) as mRNAs and the user provided sRNA sequence. The structural accessibilities for mRNAs are precomputed, while the sRNA accessibility is computed during the query (see Subsection 4.1.2).

Interactions energies are calculated for the sRNA with each mRNA of interest. In Figure 4.2 the distribution of mRNAs per genome or plasmid is shown. There is a large group of replicons that encodes fewer than 500 mRNAs, but the majority of replicons has a higher number. While being more sensitive, methods with a high time complexity are not suitable for a web server that computes interactions for a high number of mRNAs.

In every query the targets are defined by choice of target DNA molecule. That enabled us to compute the accessibilities for all the targets in advance of the target prediction and avoid the dilemma of either deciding for run time or sensitivity.

RNApex [101] (see Subsection 3.3.5) is the fastest implementation featuring a physics-based model available and allows to consider intramolecular base-pairing of both sRNA (query) and mRNA (target) sequence by including the precomputed structural accessibilities.

The adjustable parameter interaction length was set to 30 or length of the sRNA if shorter, duplex distance was set to 20, energy-threshold was set to 8. Usually only perfect complementary cis-acting RNAs have regions of interaction that are longer than 30, but they can be identified during post-processing with RNAup. RNApex is also able to return not only the minimal free energy result for a query-target combination, but also sub-optimal results. It is possible to recover distinct sub-optimal interaction sites by using duplex distance, which limits the overlap of participating bases and energy threshold, which limits the returned hits to an minimal total energy of interaction (in our case -8 kcal/mol).

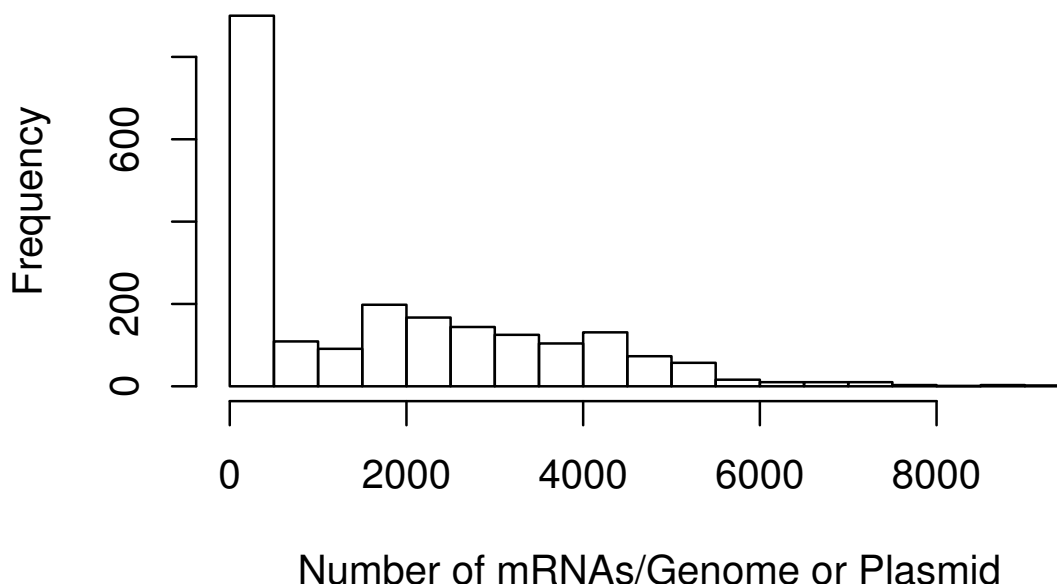


Figure 4.2: Distribution of mRNA number per genome/plasmid. A large group, containing the plasmid fraction, with fewer than 500 mRNAs encoded is shown as left-most bar. The majority of genomes has more than 1000 mRNAs, which favors methods with lower time complexity for RNA-RNA interaction studies.

The Listing 4.2 shows an example output of RNApIex as used in RNApredator, which can be retrieved for further analysis.

```
>ref_NC_009085.1_96-1492
>sRNA
(((((((.(((&)))))))))) 521,531 : 25,35 (-3.85 = -10.40 + 3.85 + 2.70)
>ref_NC_009085.1_1621-2738
>sRNA
(((((((.(((&)))))))))) 68,81 : 20,32 (-3.99 = -9.50 + 1.59 + 3.92)
```

Listing 4.2: RNApIex output showing two results from the file. The first line shows the unique locus identifier used in RNApredator, consisting of NCBI accession number identifying the genome/plasmid and the genomic coordinates corresponding with the mRNA. The second line is reserved for the sRNA identifier. The third line contains the results for the interaction prediction of the mRNA and the sRNA specified in lines 1 and 2. At the beginning of the line we find a dot bracket-string for the hybridization, followed by coordinates of the hybridization site on the hypothetical transcript, coordinates on the sRNA and the total hybridization energy modified by the opening energies derived from the structural accessibility files of mRNA and sRNA.

RNApIex output is converted into a comma separated value file and used for target evaluation.

4.3 Target evaluation

In general the reason for performing target prediction for sRNAs is to determine their biological function. sRNAs achieve their function via interaction to their targets. If the set of interaction partners is known, this simplifies the identification of the biological function significantly.

For each mRNA-sRNA interaction and additional sub-optimal results RNAplex returns a hybridization energy. Not all of the mRNAs interact with the limited number of sRNAs in the cell in sufficient manner to affect the cell. It is not possible to decide upon a fixed energy cutoff that divides the predicted hits into biologically relevant or irrelevant ones and is valid for all interactions. An alternative is to select a subset of best interactions and consider them for further target evaluation.

4.3.1 Energy-based ranking

To simplify energy based ranking the hits are sorted and the energy z-score is calculated. This information is summarized together with the dot bracket string and the coordinates of the hybridization, where the first nucleotide of the translation start is set to 0. Additionally annotations from the NCBI-genome files are retrieved and added.

Based on this preselection computationally more expensive post-processing steps can be applied, e.g. reanalyzing the hybridization with RNAup, change in regulation, or GO-term enrichment.

4.3.2 Change in regulation

One of the most interesting questions for each interaction between mRNAs and sRNAs is whether translation is upregulated or not. If the hybridization site overlaps with the ribosome binding site (RBS), then a down-regulation will be very likely the result. But if the interaction occurs in a region more distant from the RBS this question becomes non-trivial. By computing the structural accessibility of the RBS, before and after sRNA binding this process can be modeled.

RNAup [95] is called with and without constraint on the hybridization site. The adjustable parameter length of the unstructured region was set to 4, the constraint was set to begin and end of the hybridization site coordinates. Instead of just reducing the output of the accessibility change calculation to a binary up- or downregulation, accessibility with/without constraint and the resulting difference is plotted for 200 nucleotides around the transcription start of the hypothetical transcript with a R. This allows interpretation of special

cases where the RBS is partly getting more and less accessible. An example is shown in Figure 5.10.

Other mechanisms, like changes in stability or degradation are not covered by this method.

4.3.3 GO-term enrichment

As described before, sRNAs function depends on interaction with their targets. Depending on the sRNA, some will affect targets in a widespread manner to cause many minor changes, while others are focused on a certain biological pathway or process. The biological function of the second group can be better understood with an enrichment analysis of GO terms [106] (see Section 3.4) for the set of selected targets. These targets are associated with with GO-terms, each of them belonging to the GO categories (Biological Process, Molecular Function, Cellular Component).

The 20 highest enriched terms of the selected targets are returned in tabular format showing GO-ID, annotated term, total number of genes linked to this GO-ID, total number of predicted targets linked to this GO-ID, number of expected linked targets as well as the P-value are returned. An example is shown in Table 5.1.

Bacterial GO term flat-files, which are necessary for the GO term enrichment analysis were downloaded from <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes> for available species. A table to convert the ebi species IDs to the NCBI taxonomy IDs used in RNAPredator was also provided there. GO term enrichment is based upon these files and an R-script [110] relying on the TopGO [111] library.

The results of the enrichment heavily depend on the quality of the GO-term annotation (which is constantly improving), but significant representation of certain terms can be an indicator for the biological function.

5 Results - RNAPredator

The RNAPredator webserver provides automated sRNA target prediction and has been published in the NAR Webserver Issue 2011 [1].

A series of strategies has been developed for sRNA target prediction. The sRNA micC [112] and istR-1 [113] targets were identified with BLAST. A SmithWaterman recursion scoring the base pairing potential of two RNAs was used in TargetRNA [114]. Mandin et al. [115] relied on a model, where base pair stacks are scored according to the standard RNA folding energy model [84] and optimized bulge penalties.

RNAup [95, 116], IntaRNA [117] and biRNA [118] offer more general RNA-RNA interaction prediction using the RNA folding energy model, considering structural accessibilities but have higher time complexity.

RNAPredator relies on the fast dynamic programming approach, RNAPlex [105] (see Subsection 3.3.5), to predict putative targets.

In addition to a fast prediction time, RNAPredator offers additional tools to further interpret targets, like Gene Ontology (GO) enrichment analysis and a visualization of the structural accessibility change upon sRNA binding.

5.1 Motivation

As mentioned before about 80 sRNAs are annotated for the *E. coli* genome, and 140 for all species [64]. Extrapolated from *E. coli* for all bacteria contained in RNAPredator that would leave over 100000 sRNAs yet to be identified and functionally characterized. Not all discovered sRNAs have been added to databases, which further complicates the situation.

sRNA realize their effect by interaction with mRNAs, thus mRNA annotation can be used to infer the biological function of the sRNA. An automated target prediction that narrows the number of targets that have to be analyzed experimentally, saves resources and accelerates the characterization process.

RNAPredator provides automatic target prediction and tools to further evaluate the returned targets. We decided to offer this service in a manner that does not require extensive setup time, configuration or additional software, only a browser. The goal was to make RNAPredator available to as

many users as possible. A set of 1183 bacterial species consisting of 2154 genomes and plasmids is available for selection. RNAPredator is available at <http://rna.tbi.univie.ac.at/RNAPredator>.

5.2 Overview of RNAPredator functionality

At the beginning of a RNAPredator session, the user provides a sRNA sequence and selects a genome or plasmid. RNAPredator computes a prediction with RNAplex for each annotated mRNA with the sRNA. Usage of the pre-computed accessibilities yields interaction accuracies similar to more complex and computationally more costly strategies, while being at least three orders of magnitude faster than alternative methods considering target site accessibilities [1]. RNAPredator can therefore be easily applied to whole genomes.

RNAPredator returns a list of target sites sorted by the energy of interaction. For a user-defined subset of predictions an enrichment analysis of GO terms can be performed. Additionally, an possible up- or down-regulation of target translation by sRNA binding can be studied with RNAup, showing the accessibility of the ribosome binding site [116].

5.3 Pipeline and implementation

The pipeline will be described following the sequence of steps necessary to process a request, considering input, processing and output. The server side of RNAPredator is implemented in Perl 5, while most of the features on the client side, are provided by javascript. This reduces response time for smaller worksteps, like sorting of lists, that are computed directly on the client.

The data used was downloaded from NCBI (<ftp://ftp.ncbi.nih.gov/genomes/>) and prepared with the methods hypothetical transcript construction (see Subsection 4.1.1) and structural accessibility calculation (see Subsection 4.1.1).

The pipeline showing the data-flow is visualized in Figure 5.1. After the genome/plasmid has been selected by the user and the sRNA sequences (up to 5) has been entered, a URL is created where all results can be accessed. sRNA accessibilities are computed with RNAup, while the mRNA accessibilities are already precomputed.

Queuing and management of job queries is done by Sun Grid Engine 2.6u3, making the submission of multiple jobs in parallel possible. RNAPredator is hosted by Intel(R) Core(TM)2 Quad CPU Q9450 @ 2.66GHz machines running Fedora Core 12 as operating system. Each mRNA sequence is tested against the sRNA with RNAplex, considering accessibilities of both (for details see Subsection 4.2). GO enrichment analysis is based on R [110] and the

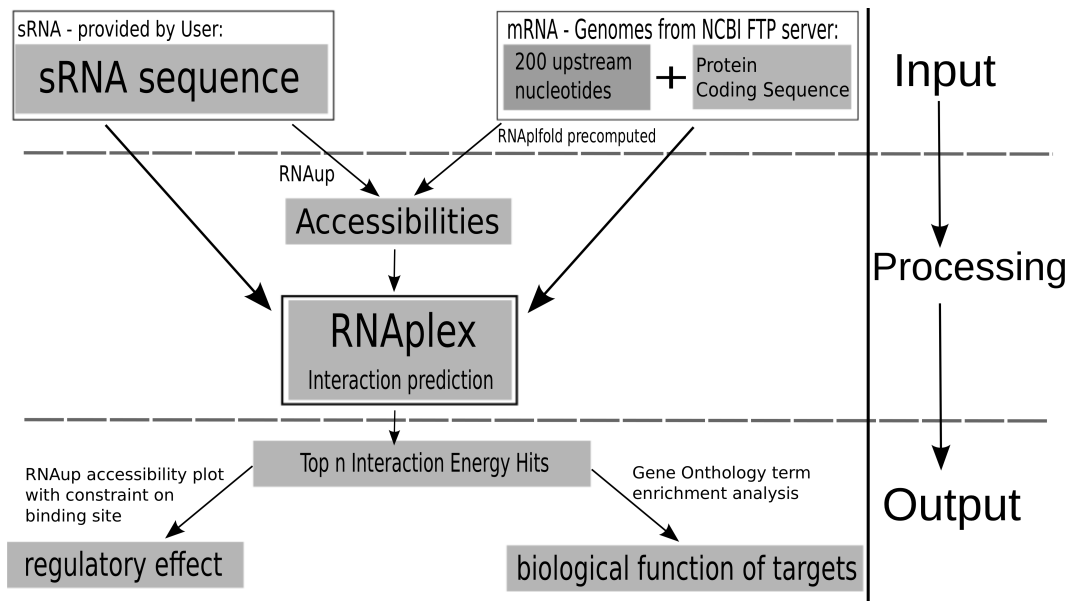


Figure 5.1: RNApredator pipeline visualization showing the program-flow from input (top) over processing (middle part) to output and postprocessing (bottom). Input consists of the sRNA sequence entered by the user and precomputed data. Processing is performed by RNApflex. The results can be used for further postprocessing (Accessibility plot, GO-term enrichment).

R library TopGO [111] (see Subsection 4.3.3). Accessibility change of the ribosome binding site is computed with RNAup and plotted with R (see Subsection 4.3.2).

5.4 Input

The first step is the choice of DNA molecule of interest. For the regulatory role some sRNAs it may be necessary to consider also plasmid encoded mRNAs alongside genome encoded ones. RNApredator offers 3 ways to select the genome and plasmid combination of interest.

The first page being displayed when browsing to <http://rna.tbi.univie.ac.at/RNApredator> is the *Target Search* menu. This is one of 4 selections offered by the main menu shown in Figure 5.2:

Introduction and *Help* provide background and usage information, while *Available Genomes* and *Target Search* offer the functionality of the web-service.

Target Search is the starting point for users who already know either the NCBI accession number of the genome or plasmid of interest or the NCBI taxonomy ID associated with a group of DNA molecules.




Figure 5.2: Navigation bar of RNAPredator: *Introduction* gives a short description of the webserver. *Available Genomes* offers the tree view and the search function of provided genomes. *Target Search* provides the prediction service and *Help* contains a How-To.


Alternatively the *Available Genomes* selection page can be used offering a phylogenetic tree and a search function for provided genomes/plasmids and species.

A phylogenetic tree can be used to browse for the replicon of interest. The tree is based on taxonomic data from NCBI (<ftp://ftp.ncbi.nih.gov/pub/taxonomy>) and implemented with the jquery js-tree plug-in.

Click on a Species or Genome Node:

Taxonomic Node: 

Species Node: 

Genome Node: 

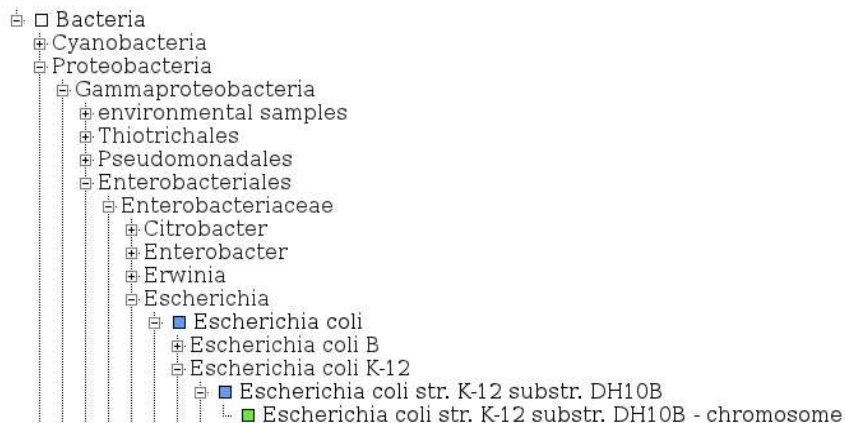


Figure 5.3: Phylogenetic tree showing available genomes grouped into phylum/class/order/family/genus/species. Three types of nodes can be found in the tree: taxonomic node (grey) - containing no genomes, species node (blue) - containing at least one genome/plasmid child node, genome node (green) - containing a single genome/plasmid.

The phylogenetic tree (see Figure 5.3) is grouped into phylum/class/order/family/genus/species nodes. Taxonomic nodes do not contain genomes but can have species or genome nodes as child nodes. Clicking a genome node selects a specific genome or plasmid for target search. Clicking a species node selects all direct child nodes of type genome node for target search. This makes it possible to search against a genome and several plasmids.

Search for a Genome or Plasmid:

Search

e.g.: *Escherichia coli str. K-12 substr. MG1655 chromosome*
 You will be redirected to target search upon selection

Figure 5.4: Genome search field with suggestion box (not shown).

A more direct way is the search box (see Figure 5.4) at the top of the *Available Genomes* page, that features a suggest function and displays a search result page (see Figure 5.5) on clicking the search button.

Query: "Escherichia coli str"

[Back to Available Genomes](#)

Designation	Accession Number	Select for Target Search
Escherichia coli str. K-12 substr. DH10B chromosome	NC_010473	Link
Escherichia coli str. K-12 substr. W3110 chromosome	AC_000091	Link
Escherichia coli str. K-12 substr. MG1655 chromosome	NC_000913	Link

Figure 5.5: Genome search results displaying all entries containing the string supplied on the *Available Genomes* page. Using the corresponding link selects for *Target Search* and redirects there.

Upon selection the browser returns to *Target Search* with the corresponding IDs already set.

Once the selection of genome or plasmid has been made three steps follow. The first is a check if the correct genomes and plasmids have been selected (see 2. *Confirm Genome* in Figure 5.6).

Then the sRNA sequence is typed in the provided text box or an File in fasta format containing one or up to 5 sRNA sequences is uploaded (see 3. *Enter sRNA-Sequence and Submit* in Figure 5.6). All lower-case letters will be replaced by their upper-case equivalent and T with U. If the format is incorrect the user is notified by a warning message. Optionally it is possible to provide an email-address for notification on completion of the prediction. Pressing the predict button submits the input to the queue for prediction.

During the computation the progress is displayed, as shown in Figure 5.7:

1. Select Genome:	2. Confirm Genome:	3. Enter sRNA-Sequence and Submit
NCBI-Accession Number <input type="text" value="NC_000913"/>	NCBI Accession Number: NC_000913	<input type="text" value="GAAAGACGCGCAUUUUGUUAUCAUCAUCCUGAAUUCAGAG
AUGAAAUUUUGGCCACUCACGAGUGGCCUUUUUCUUUU"/>
OR	Name: Escherichia coli str. K-12 substr. MG1655	OR
NCBI-Taxonomy ID <input type="text"/>	Tax-id: 511145	Upload a fasta-file: <input type="text"/> <input type="button" value="Browse..."/>
OR	Replicon: chromosome	Please provide email-address for notification: <input type="text"/>
Select from Available Genomes		(optional)
		<input type="button" value="Predict"/> <input type="button" value="Back"/> <input type="button" value="Reset"/>

Figure 5.6: Steps required for target prediction. The example shows the input for Escherichia coli. K-12 substr. MG1655 and MicA sRNA. Alternatively the Taxonomy ID field or *Available Genomes* could be used to define the genomes and plasmids of interest. sRNA submit is possible by entering a sequence or uploading of a file in fasta format.

This can take some minutes.

Wait for calculation to finish or return [here](#) later.

RNA	Queuing Status	sRNA accessibility	RNAplex Interaction Prediction	Parsing Output	Result Page Link
1	done	done	done	done	Link
2	done	done	done	done	Link

Figure 5.7: Prediction progress is monitored and reported back to browser. The completed results are linked and the overview page can be bookmarked for later retrieval.

5.5 Output

The default output is a list showing the top 100 hits sorted by hybridization energy. This list also contains suboptimal hits as distinct entries.

The result page offers two filtering options to display only the top 25, 50, 75, 100, 500 or all interactions by energy and / or limit the displayed entries to a specific region of interaction on the mRNA, e.g. the 5'-UTR. The result list, which can be downloaded as comma separated value file, contains 14 columns. The left-most selection box tags entries for postprocessing on click of the button Post-Process. This is followed by interaction energy rank, energy of interaction [kcal/mol], z-score of interactions energies, a dot-bracket string defining the region of hybridization, start and end on the hypothetical mRNA, hybridization coordinates on the sRNA, gene annotation taken from NCBI genome files, NCBI locus tag of the gene, DNA strand the gene is encoded on, genomic coordinates the hypothetical transcript is transcribed from, NCBI accession number of the genome or plasmid, type of the replicon (genome/plasmid). The list can be sorted by a JQuery script by clicking on

Interaction1											
Energy [kJ/mol]	z-Score	Interaction [dot-bracket]	mRNA [Start]	mRNA [End]	sRNA [Start]	sRNA [End]	Gene Annotation	Locus Tag	Genomic Coordinates	Accession	Replicon
-13.94	-1.69	(((((((((((&)))))))))))))	-18	-5	8	22	"outer membrane protein A (3a,II*;G;d)"	b0957	c1019476-1018236	NC_000913	chromosome
Associated GO-terms	GO:0000746 GO:0005198 GO:0005515 GO:0005886 GO:0006810 GO:0006811 GO:0009279 GO:0009597 GO:0015288 GO:0016020 GO:0016021 GO:0019867 GO:0046718 GO:0046930										
download: mRNA sequence download: sRNA sequence Accessibility Plot: Calculate (new window) RNAup Webserver: Submit (new window)	Detailed Interaction(as ASCII): mRNA: 5' - UGAUAACGAG-GCG- 3' sRNA: 3' - ACJAUUGUUACGC- 5'										

Figure 5.9: *E. coli str. K-12 substr. MG1655* sRNA MicA interaction with the mRNA OmpA is shown, which is one of the experimentally verified targets. Additionally to the information displayed in the result table a detailed sequence representation of the interaction and the GO-terms associated to the mRNA are shown. Additional postprocessing options are located on the left side of the entry and include retrieval of the mRNA or sRNA sequence, calculation of the accessibility plot or recomputation of the interaction with the RNAup webserver (<http://http://rna.tbi.univie.ac.at/cgi-bin/RNAup.cgi>).

for all genes. If terms are enriched significantly for a specific molecular function, biological process or cellular component, this could give an indication about the regulatory role of the sRNA.

The GO-term statistics (see Table 5.1), were enriched for the 25 best MicA targets from *E. coli str. K-12 substr. MG1655*. The top 4 hits are on the opposite strand of the sRNA and were not considered.

The enrichment statistics shows the 20 most significant terms per category. The columns from left to right show: the GO-term ID, the human readable annotation, how many such genes are annotated for the species, how many genes in the selection have to be associated with that specific GO-term for significant enrichment, the expect value describing the number of this GO-term one can see when searching by chance and the weight 01 p-value describing the significance of the enrichment.

The cellular component enrichment table (see Table 5.1) shows GO-terms for periplasmatic space, extracellular region, integral to membrane and cell outer membrane. This would indicate an connection between the biological function of the sRNA and the outer membrane. For a novel sRNA this could add a valueable perspective the the characterisation process. In case of MicA, which is known to regulate outer membrane proteins [119], it just confirms what is already known.

GO.ID	Term	Annotated	Significant	Expected	Weight01 p-value
GO:0042597	periplasmic space	180	4	1.01	0.022
GO:0032155	cell division site part	12	1	0.07	0.066
GO:0005576	extracellular region	15	1	0.08	0.081
GO:0005694	chromosome	19	1	0.11	0.102
GO:0016021	integral to membrane	912	7	5.14	0.217
GO:0030288	outer membrane-bounded periplasmic space	61	1	0.34	0.295
GO:0009279	cell outer membrane	82	1	0.46	0.376
GO:0005886	plasma membrane	1073	7	6.04	0.398
GO:0005737	cytoplasm	822	1	4.63	0.997
GO:0043228	non-membrane-bounded organelle	116	1	0.65	1.000
GO:0043229	intracellular organelle	135	1	0.76	1.000
GO:0043232	intracellular non-membrane-bounded organ...	116	1	0.65	1.000
GO:0005622	intracellular	1024	2	5.77	1.000
GO:0005623	cell	2302	13	12.97	1.000
GO:0044424	intracellular part	898	2	5.06	1.000
GO:0005575	cellular_component	2308	13	13	1.000
GO:0044425	membrane part	956	7	5.38	1.000
GO:0044462	external encapsulating structure part	142	2	0.8	1.000
GO:0031224	intrinsic to membrane	916	7	5.16	1.000
GO:0044464	cell part	2302	13	12.97	1.000

Table 5.1: Enrichment statistics for the twenty most significant GO-terms for the category Cellular Component. The enrichment was performed for the 25 best hits of MicA in *E. coli str. K-12 substr. MG1655*. The top 4 hits are on the opposite strand of the sRNA and were not considered. GO-terms for the periphery of the cell are the most significant ones.

The quality of the GO-term annotations is of major importance to the meaningfulness of this approach.

A further source of information is the influence of sRNA binding on the accessibility of the ribosome binding site. An increase of accessibility could indicate an upregulation of translation, while a decrease could indicate a downregulation. The example shown in Figure 5.10 is again from *E. coli str. K-12 substr. MG1655* and shows the plot for MicA and OmpA.

The accessibility plot (see Figure 5.10) shows a decrease of accessibility in the ribosome binding site. MicA is known to downregulate OmpA, by binding in this region and additionally by facilitating degradation [119], which is consistent with this prediction.

5.6 Benchmark

Thirty experimentally verified interactions from literature were used to compare RNApredator [1] to TargetRNA [114]. The RNAup webserver was not considered as it is not designed for genome-wide target predictions. Only interactions predicted to hybridize to hypothetical transcript coordinates between the interval -150 up to 100 and -30 up to 20 relative to the start codon were included. TargetRNA was set to hybridization length 1, G-U base pairs allowed and 100 as p-value cutoff. The thermodynamic scoring of TargetRNA

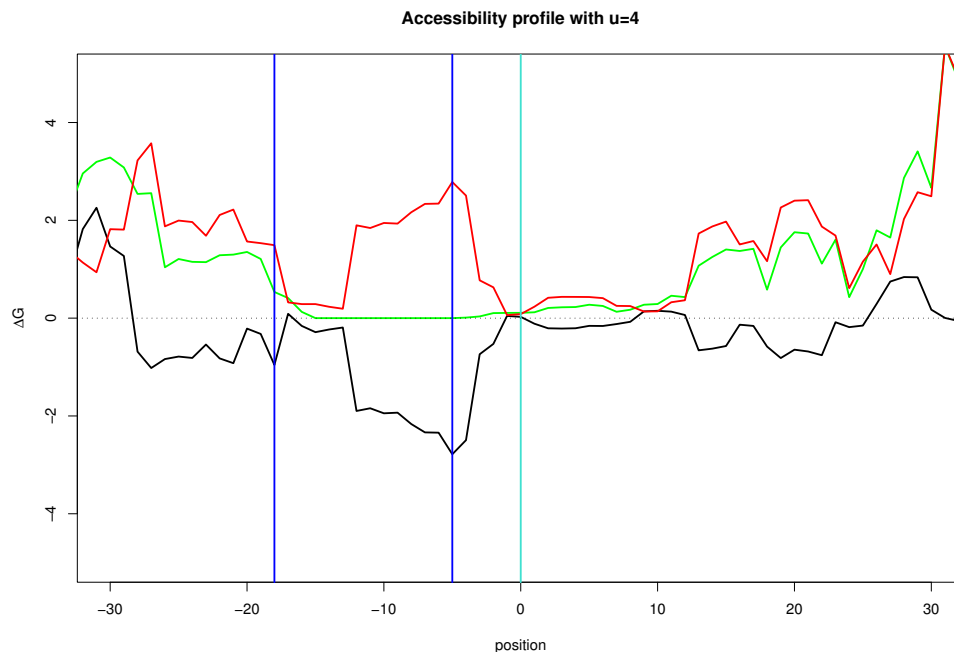


Figure 5.10: Accessibility profiles for *E. coli str. K-12 substr. MG1655* MicA binding OmpA. The red line shows accessibility before binding, the green line shows accessibility after binding, the black line indicates the difference of both. The blue lines show the hybridization site and the cyan line indicates the start codon. The predicted hybridization occurs directly in the ribosome binding site.

was not returning results and could therefore not be considered. Table 5.2 is featured in the RNApredator publication [1] and included for completeness. It compares the rank of experimentally verified targets found by prediction. TargetRNA only returns at most the 100 best interactions, therefore some experimentally verified interactions could not be found in the result. The only interactions not found in RNApredator are those that were predicted to hybridize outside the used constraint (-150 to 100 and -30 to 20). RNApredator ranks in 73% of the interactions with experimentally verified targets considered in Table 5.2 better than TargetRNA [1].

5.7 Usage statistics

In Figure 5.11 the usage of RNApredator since publication on 16th of June 2011 in Nucleic Acid Research Web Server Issue 2011 [1]. Between publication and 31st of October 2011 1902 requests were submitted, as shown in Figure 5.11.

Genome	Species	sRNA	mRNA	Gene	TargetRNA	RNApredator
NC_000964	B.s.	FsrA	sdhC	BSU28450	NF(NF)	153(83)
NC_011601	E.c.O	OmrA	ompR	b3405	NF(NF)	436(49)
NC_011601	E.c.O	OmrA	ompT	b0565	NF(NF)	712(93)
NC_011601	E.c.O	OmrB	ompR	b3405	NF(31)	312(39)
NC_011601	E.c.O	OmrB	ompT	b0565	NF(NF)	210(13)
NC_000913	E.c.K.	CyaR	ompX	b0814	NF(NF)	495(86)
NC_000913	E.c.K.	CyaR	yqaE	b2666	NF(NF)	541(97)
NC_000913	E.c.K.	DsrA	hns	b1237	52(6)	8(4)
NC_000913	E.c.K.	FnrS	metE	b3829	5(8)	120(37)
NC_000913	E.c.K.	FnrS	sodB	b1656	24(21)	615(192)
NC_000913	E.c.K.	GcvB	cycA	b4208	37(5)	41(10)
NC_000913	E.c.K.	IstR	tisB	b4405	2(NF)	NF(NF)
NC_000913	E.c.K.	MicA	phoP	b1130	80(23)	57(10)
NC_000913	E.c.K.	MicC	ompC	b2215	2(5)	2(2)
NC_000913	E.c.K.	MicF	ompF	b0929	43(5)	2(2)
NC_000913	E.c.K.	OmrA	gntP	b4321	NF(NF)	79(17)
NC_000913	E.c.K.	OmrB	csgD	b1040	50(NF)	2(NF)
NC_000913	E.c.K.	RseX	ompC	b2215	98(NF)	504(238)
NC_000913	E.c.K.	RyhB	iscS	b2530	NF(NF)	123(30)
NC_000913	E.c.K.	RyhB	sodB	b1656	24(21)	184(52)
NC_000913	E.c.K.	SgrS	ptsG	b1101	NF(NF)	5(1)
NC_003210	L.m.	LhrA	lmo085	lmo0850	NF(NF)	31(NF)
NC_002505	V.c.	MicX	vca0620	vca0620	NF(34)	48(7)
NC_002505	V.c.	Qrr1	luxO	vca1021	NF(NF)	196(44)
NC_002505	V.c.	Qrr1	vca0939	vca0939	NF(NF)	5(NF)
NC_002505	V.c.	Qrr2	luxO	vca0620	NF(NF)	12(NF)
NC_002505	V.c.	Qrr2	vca0939	vca0939	NF(NF)	3(NF)
NC_002505	V.c.	Qrr3	vca0939	vca0939	NF(NF)	4(NF)
NC_002505	V.c.	Qrr4	vca0939	vca0939	NF(NF)	4(NF)
NC_002505	V.c.	VrrA	tcpA	vca0838	35(NF)	246(71)

Table 5.2: Table shows a benchmark of RNApredator against TargetRNA. The columns from left to right: NCBI accession number, species abbreviation (B.s.=Bacillus subtilis subsp. subtilis str. 168, E.c.O=Escherichia coli O127:H6 str. E2348/69, E.c.K=Escherichia coli str. K-12 substr. MG1655, L.m.=Listeria monocytogenes EGD-e and V.c.=Vibrio cholerae O1 biovar El Tor str. N16961), sRNA NCBI gene tag, mRNA NCBI gene tag, NCBI locus tag, interaction rank TargetRNA, interaction rank RNApredator. The first number provided in the rank column corresponds to predictions only considering hybridizations between coordinates -30 and 20, while the number in brackets corresponds to predictions constraint to hybridizations between -150 and 100, relative to the start codon. NF is abbreviated for not found. [1] by permission of Oxford University Press.

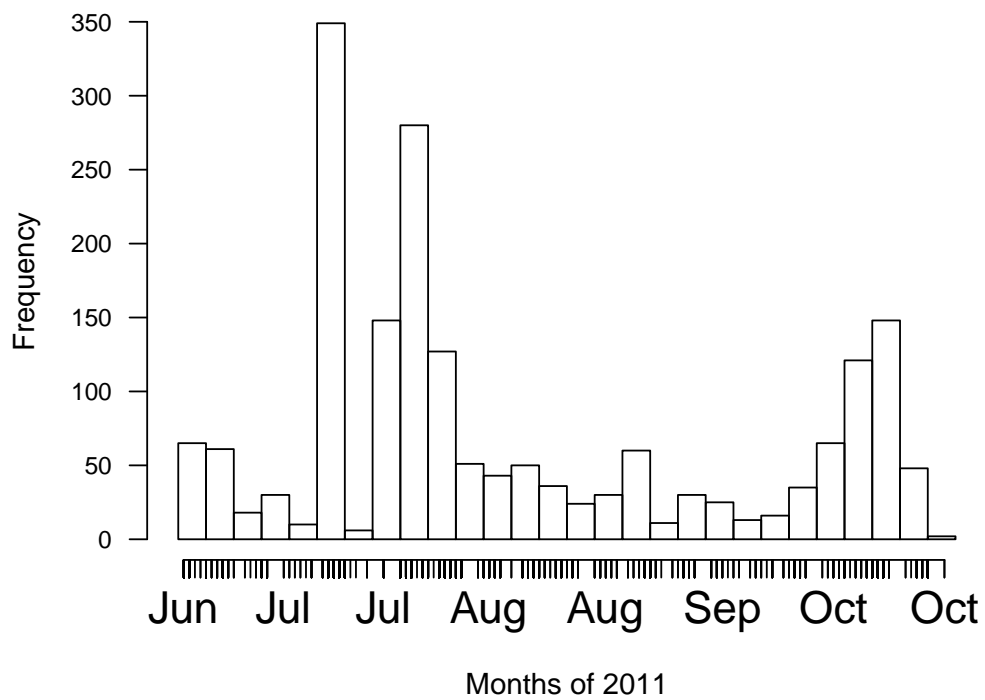


Figure 5.11: Usage-statistics of RNApredator, counted by submitted requests, not considering additional post processing requests.

6 Discussion and outlook

A substantial number of sRNAs from sequenced species is not yet functionally characterized (see extrapolation in Section 5.1). The experimental effort needed to meet this challenge is considerable, nevertheless it is essential to understand this bacterial mechanism of regulation. Beside the scientific progress, applications in biotechnology, medicine, ecology, etc. could arise from it.

While current sRNA interaction prediction *in silico* cannot replace *in vivo* experiments, it can simplify the experimental approach and save both time and resources by providing a set of candidate mRNAs.

RNApredator is a freely available web-tool providing sRNA target prediction *in silico*. It does not require a complicated set-up procedure, only a web-browser is mandatory.

The predictions made by RNApredator are saving at least three orders of magnitude of computation time, while reaching similar accuracy compared to more complex methods [1]. This is due to precomputed structural accessibilities that represent intramolecular base-pairing of the molecules prior to interaction. Accessibilities are calculated with RNAplfold [91], while target prediction is performed by RNAplex [105].

The resulting interaction list provides a first selection for experimental analysis. This is complemented by the available post-processing tools, which are unique for RNApredator. GO-term enrichment can give a first clue about common characteristics of the target gene-set, while the accessibility change plot yields information about the regulation of individual genes. GO-term enrichment relies on the Top-GO R-library [111]. The accessibility-plot is computed using RNAup [116].

While RNApredator results are acceptable in many cases, there are factors like protein interactions (e.g Hfq in *E. coli*) that distort the results. Another factor is temperature, that strongly influence the computed binding energies and should be considered for bacteria that dwell in corresponding environments.

These challenges suggest what can be done to further improve predictions.

Including RNA-Protein interactions into the prediction that allow to consider e.g. RNA chaperons would broaden the spectrum of sRNAs RNApredator can be applied to successfully. In its simplest form a post-processing step could be added that scans for Hfq-binding motifs on target mRNAs, that possibly

indicates Hfq involvement.

Another ambitious enhancement would be to consider conservation of sRNAs and mRNAs over several species. The required algorithm is already implemented into the newest version of RNAplex [105] (included in the Vienna RNA Package 2.0). Sequence alignments of both homologous sRNAs and mRNAs are used as input. This approach heavily depends on alignment quality and on the availability of homologous sRNAs to use. This could be included by either uploading a set of aligned sRNAs, or instead a automated homology search that is conducted by the server.

Using temperature as parameter is already implemented in the Vienna RNA Package, but accessibilities are precalculated and the temperature would have to be considered at that step. A solution could be to precompute several sets of accessibilities for different temperatures. If regulatory activities of sRNAs change with temperature it could be interesting to compare the results of predictions for different temperatures.

Preferably the server should use real transcript sequences from e.g. RNA sequencing instead of hypothetical transcripts. Until these are available, species specific parameters could be used to increase hypothetical transcript quality.

The consideration of biological context data can generally be used to optimize prediction results. Known expression patterns could decrease the number of possible sRNA targets to those transcripts that are present at the same time as the sRNA. This could be realized by adding a list of mRNA IDs, which are expressed simultaneously, to the request.

Other possible improvements can be made concerning the architecture of the pipeline.

RNAplex and RNAup are limited to predict one hybridization site only. That precludes interactions like OxyS with *fhlA*, which feature 2 kissing hairpin interactions. Heuristics could be added that group them together.

Multi-core CPU architectures become more and more common. Prediction of each sRNA-mRNA interaction is independent, which enables parallelisation of the predictions and runtime reduction.

Also the post-processing tools can be further improved and extended.

While GO-term enrichment is already implemented, also Kegg [120] term enrichment that could be included and improve the information provided for selected targets.

sRNA and mRNA interactions are concentration dependend processes. If concentration patterns of sRNAs and mRNAs are available, they could be used to

produce quantitative results showing how many mRNAs are bound/unbound. As mentioned in the biological background [63], other ncRNAs exist in archaea and eukarya. Predictions in eukarya are further complicated by RNA export, RNA-protein interactions and the increased complexity of the system. Moreover several specialized tools for miRNA target prediction in eukarya are already available.

Some genomes of archaea are already available in RNAPredator, but a tool featuring all archaea could be useful. RNAPredator could serve as template for this pipeline.

Bibliography

- [1] Florian Eggenhofer, Hakim Tafer, Peter F. Stadler, and Ivo L. Hofacker. RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acid Research*, (Web Server Issue 2011):6, 2011.
- [2] W Gilbert. Origin of life: The rna world. *Nature*, 319:618, February 1986.
- [3] L.E. Orgel and F.H.C. Crick. Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607, 1980.
- [4] J.B. Sumner. The isolation and crystallization of the enzyme urease. *Journal of Biological Chemistry*, 69(2):435, 1926.
- [5] M. Smith F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe. The nucleotide sequence of bacteriophage X174. *Nature*, (265):687 – 695, 1977.
- [6] W Fiers, R Contreras, F Duerinck, and G Haegeman. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, (260):500 – 507, 1976.
- [7] Shoshy Altuvia. Identification of bacterial small non-coding RNAs: experimental approaches. *Current opinion in microbiology*, 10(3):257–61, June 2007.
- [8] F Crick. Central dogma of molecular biology. *Nature*, 227:561–563, August 1970.
- [9] B Alberts, A Johnson, J Lewis, and et al. *Molecular Biology of the Cell. 4th edition*. New York: Garland Science, 2002.
- [10] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.
- [11] James C Wang. Helical repeat of DNA in solution *Biochemistry* : Wang. *Biochemistry*, 76(1):200–203, 1979.
- [12] Timothy J Richmond and Curt a Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–50, May 2003.
- [13] A J Zaug and T R Cech. In vitro splicing of the ribosomal RNA precursor in nuclei of Tetrahymena. *Cell*, 19(2):331–338, 1980.

- [14] C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–857, 1983.
- [15] K Kruger, P J Grabowski, A J Zaug, J Sands, D E Gottschling, and T R Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, 31(1):147–157, 1982.
- [16] Geometric nomenclature and classification of RNA base pairs. *RNA (New York, N.Y.)*, 7(4):499–512, April 2001.
- [17] Gabriele Varani and William H McClain. EMBO reports The G U wobble base pair diverse biological systems. *Molecular Biology*, 1(1):18–23, 2000.
- [18] Donna K Hendrix, Steven E Brenner, and Stephen R Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Quarterly reviews of biophysics*, 38(3):221–43, August 2005.
- [19] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic acids research*, 34(2):564–74, January 2006.
- [20] Thomas Hermann and D.J. Patel. RNA bulges as architectural and recognition motifs. *Structure*, 8(3):R47–R54, 2000.
- [21] H. H. Gan. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Research*, 31(11):2926–2943, June 2003.
- [22] K Gruenberger. *3D-Model for coarse grained Structure Prediction of RNA*. PhD thesis, University of Vienna, 2002.
- [23] I Tinoco and C Bustamante. How RNA folds. *Journal of molecular biology*, 293(2):271–81, October 1999.
- [24] Ivo L. Hofacker and Peter F. Stadler. RNA secondary structures. In Thomas Lengauer, editor, *Bioinformatics: From Genomes to Therapies*, volume 1, pages 439–489. Wiley-VCH, Weinheim, Germany, 2007.
- [25] H F Noller. Structure of ribosomal RNA. *Annual review of biochemistry*, 53:119–62, January 1984.
- [26] Stephan Bernhart. Sadsat - a knowledge based potential for protein folding. Master’s thesis, University of Vienna, 2003.
- [27] L. Pauling, R.B. Corey, and H.R. Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 37(4):205, 1951.

- [28] S T Rao and M G Rossmann. Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, 76(2):241–256, 1973.
- [29] M H Cordes, a R Davidson, and R T Sauer. Sequence space, folding and protein design. *Current opinion in structural biology*, 6(1):3–10, February 1996.
- [30] I M Klotz, N R Langerman, and D W Darnall. Quaternary structure of proteins. *Annual review of biochemistry*, 39:25–62, January 1970.
- [31] Douglas F Browning and Stephen J Busby. The regulation of bacterial transcription initiation. *Nature reviews. Microbiology*, 2(1):57–65, January 2004.
- [32] E V Koonin. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *The Journal of general virology*, 72 (Pt 9):2197–206, September 1991.
- [33] Joshua W Shaevitz, Elio A Abbondanzieri, Robert Landick, and Steven M Block. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature*, 426(6967):684–687, 2003.
- [34] R T Libby and J A Gallant. Phosphorolytic error correction during transcription. *Molecular Microbiology*, 12(1):121–129, 1994.
- [35] L. F. Liu. Supercoiling of the DNA Template during Transcription. *Proceedings of the National Academy of Sciences*, 84(20):7024–7027, October 1987.
- [36] John P Richardson. Rho-dependent termination and ATPases in transcript termination. *Biochimica et biophysica acta*, 1577(2):251–260, September 2002.
- [37] P.J. Farnham and Terry Platt. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro. *Nucleic acids research*, 9(3):563, 1981.
- [38] AJ Carpousis. The Escherichia coli RNA degradosome: structure, function and relationship to other ribonucleolytic multienzyme complexes. *Biochemical Society Transactions*, 30(2):1, 2002.
- [39] a Simonetti, S Marzi, L Jenner, a Myasnikov, P Romby, G Yusupova, B P Klaholz, and M Yusupov. A structural view of translation initiation in bacteria. *Cellular and molecular life sciences : CMLS*, 66(3):423–36, February 2009.
- [40] M M Yusupov, G Z Yusupova, a Baucom, K Lieberman, T N Earnest, J H Cate, and H F Noller. Crystal structure of the ribosome at 5.5 Å resolution. *Science (New York, N.Y.)*, 292(5518):883–96, May 2001.

- [41] M Sprinzl, C Horn, M Brown, a Ioudovitch, and S Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, 26(1):148–53, January 1998.
- [42] Naglis Malys and John E G McCarthy. Translation initiation: variations in the mechanism can be anticipated. *Cellular and molecular life sciences : CMLS*, 68(6):991–1003, March 2011.
- [43] J. Shine and L. Dalgarno. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Sciences*, 71(4):1342, 1974.
- [44] M Nirenberg, P Leder, and M Bernfield. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the*, 53(1962):1161–1168, 1965.
- [45] S Osawa, T H Jukes, K Watanabe, and a Muto. Recent evidence for evolution of the genetic code. *Microbiological reviews*, 56(1):229–64, March 1992.
- [46] Yan Zhang, Pavel V Baranov, John F Atkins, and Vadim N Gladyshev. Pyrrolysine and selenocysteine use dissimilar decoding strategies. *The Journal of biological chemistry*, 280(21):20740–51, May 2005.
- [47] I V Boni, D M Isaeva, M L Musychenko, and N V Tzareva. Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic acids research*, 19(1):155–62, January 1991.
- [48] K D Dahlquist and J D Puglisi. Interaction of translation initiation factor IF1 with the E. coli ribosomal A site. *Journal of molecular biology*, 299(1):1–15, May 2000.
- [49] RM Sundari, EA Stringer, LH Schulman, and U. Maitra. Interaction of bacterial initiation factor 2 with initiator tRNA. *Journal of Biological Chemistry*, 251(11):3338, 1976.
- [50] A Dallas and H F Noller. Interaction of translation initiation factor 3 with the 30S ribosomal subunit. *Molecular cell*, 8(4):855–64, October 2001.
- [51] Bill Chang, Saman Halgamuge, and Sen-Lin Tang. Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene*, 373:90–9, May 2006.
- [52] I V Boni, V S Artamonova, N V Tzareva, and M Dreyfus. Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1. *The EMBO journal*, 20(15):4222–32, August 2001.

- [53] Isabella Moll, Go Hirokawa, Michael C Kiel, Akira Kaji, and Udo Bläsi. Translation initiation with 70S ribosomes: an alternative pathway for leaderless mRNAs. *Nucleic acids research*, 32(11):3354–63, January 2004.
- [54] S Grill, C O Gualerzi, P Londei, and U Bläsi. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *The EMBO journal*, 19(15):4101–10, August 2000.
- [55] Isabella Moll, Sonja Grill, Claudio O Gualerzi, and Udo Bläsi. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Molecular microbiology*, 43(1):239–46, January 2002.
- [56] V Ramakrishnan. Ribosome structure and the mechanism of translation. *Cell*, 108(4):557–572, 2002.
- [57] Anatoly P. Potapov. A stereospecific mechanism for the aminoacyl-tRNA selection at the ribosome. *FEBS Letters*, 146(1):5–8, September 1982.
- [58] J.J. Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences*, 71(10):4135, 1974.
- [59] Marcus E. Peter, Christian O.a. Reiser, Norbert K. Schirmer, Thomas Kiefhaber, Günther Ott, Norbert W. Grillenbeck, and Mathias Sprinzl. Interaction of the isolated domain II/III of *Thermus thermophilus* elongation factor Tu with the nucleotide exchange factor EF-Ts. *Nucleic Acids Research*, 18(23):6889–6893, 1990.
- [60] a V Zavialov, R H Buckingham, and M Ehrenberg. A posttermination ribosomal complex is the guanine nucleotide exchange factor for peptide release factor RF3. *Cell*, 107(1):115–24, October 2001.
- [61] A Kaji, M C Kiel, G Hirokawa, A R Muto, Y Inokuchi, and H Kaji. The fourth step of protein synthesis: disassembly of the posttermination complex is catalyzed by elongation factor G and ribosome recycling factor, a near-perfect mimic of tRNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 66:515–529, 2001.
- [62] D A Dougan, D Micevski, and K N Truscott. The N-end rule pathway: From recognition by N-recognins, to destruction by AAA+proteases. *Biochimica et biophysica acta*, July 2011.
- [63] Athanasius F Bompfünowerer, Christoph Flamm, Claudia Fried, Guido Fritsch, Ivo L Hofacker, Jörg Lehmann, Kristin Missal, Axel Mosig, Bettina Müller, Sonja J Prohaska, Bärbel M R Stadler, Peter F Stadler, Andrea Tanzer, Stefan Washietl, and Christina Witwer. Evolutionary patterns of non-coding RNAs. *Theory in biosciences = Theorie in den Biowissenschaften*, 123(4):301–69, April 2005.

- [64] Francis Repoila and Fabien Darfeuille. Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biology of the cell*, 101(2):117–31, February 2009.
- [65] Gisela Storz, Shoshy Altuvia, and Karen M Wassarman. An abundance of RNA regulators. *Annual review of biochemistry*, 74:199–217, January 2005.
- [66] Jörgen Johansson and Pascale Cossart. RNA-mediated control of virulence gene expression in bacterial pathogens. *Trends in Microbiology*, 11(6):280–285, June 2003.
- [67] Matthias Gimpel, Nadja Heidrich, Ulrike Mäder, Hans Krügel, and Sabine Brantl. A dual-function sRNA from *B. subtilis*: SR1 acts as a peptide encoding mRNA on the gapA operon. *Molecular Microbiology*, 76(4):990–1009, 2010.
- [68] Susan Gottesman. The small RNA regulators of *Escherichia coli*: roles and mechanisms*. *Annual review of microbiology*, 58:303–28, January 2004.
- [69] Falk Kalamorz, Birte Reichenbach, Walter März, Bodo Rak, and Boris Görke. Feedback control of glucosamine-6-phosphate synthase GlmS expression depends on the small RNA GlmZ and involves the novel protein YhbJ in *Escherichia coli*. *Molecular microbiology*, 65(6):1518–33, September 2007.
- [70] Thorleif Møller, Thomas Franch, Peter Højrup, Douglas R Keene, Hans Peter Bächinger, Richard G Brennan, and Poul Valentin-Hansen. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Molecular cell*, 9(1):23–30, January 2002.
- [71] Aixia Zhang, Karen M. Wassarman, Carsten Rosenow, Brian C. Tjaden, Gisela Storz, and Susan Gottesman. Global analysis of small RNA and mRNA targets of Hfq. *Molecular Microbiology*, 50(4):1111–1124, October 2003.
- [72] R.A. Lease and M Belfort. A trans-acting RNA as a control switch in *Escherichia coli*: DsrA modulates function by forming alternative structures. *Proceedings of the National Academy of Sciences*, 97(18):9919, August 2000.
- [73] J.A. Opdyke, J.G. Kang, and Gisela Storz. GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. *Journal of bacteriology*, 186(20):6698, 2004.
- [74] S Altuvia, a Zhang, L Argaman, a Tiwari, and G Storz. The *Escherichia coli* OxyS regulatory RNA represses fhfA translation by blocking ribosome binding. *The EMBO journal*, 17(20):6069–75, October 1998.

- [75] Cynthia M Sharma, Fabien Darfeuille, Titia H Plantinga, and Jörg Vogel. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes & development*, 21(21):2804–17, November 2007.
- [76] Nadja Heidrich, Isabella Moll, and Sabine Brantl. In vitro analysis of the interaction between the small RNA SR1 and its primary target ahrC mRNA. *Nucleic acids research*, 35(13):4331–46, January 2007.
- [77] Teppei Morita, Yukari Mochizuki, and Hiroji Aiba. Translational repression is sufficient for gene silencing by bacterial small noncoding RNAs in the absence of mRNA destruction. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):4858–63, March 2006.
- [78] Amy T Cavanagh, Andrew D Klocko, Xiaochun Liu, and Karen M Wasarman. Promoter specificity for 6S RNA regulation of transcription is determined by core promoter sequences and competition for region 4.2 of sigma70. *Molecular microbiology*, 67(6):1242–56, March 2008.
- [79] Josef Leydold and P.F. Stadler. Minimal cycle bases of outerplanar graphs. *Elec. J. Comb*, 5:R16, 1998.
- [80] S Louise-May, P Auffinger, and E Westhof. Calculations of nucleic acid conformations. *Current opinion in structural biology*, 6(3):289–98, June 1996.
- [81] Gary M. Studnicka, Georgia M. Rahn, Ian W. Cummings, and Winston A. Salser. Computer method for predicting the secondary structure of single-stranded rna. *Nucleic Acids Research*, 5(9):3365–3388, 1978.
- [82] MS Waterman. Secondary structure of single-stranded nucleic acids. *Adv. math. suppl. studies*, 1, 1978.
- [83] R Nussinov and a B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309–13, November 1980.
- [84] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133, 1981.
- [85] Athanasius F Bompfünowerer, Rolf Backofen, Stephan H Bernhart, Jana Hertel, Ivo L Hofacker, Peter F Stadler, and Sebastian Will. Variations on RNA folding and alignment: lessons from Benasque. *Journal of mathematical biology*, 56(1-2):129–44, January 2008.

- [86] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie Chemical Monthly*, 125(2):167–188, February 1994.
- [87] D H Mathews, J Sabina, M Zuker, and D H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology*, 288(5):911–40, May 1999.
- [88] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–92, May 2004.
- [89] S Wuchty, W Fontana, I L Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, February 1999.
- [90] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.
- [91] Stephan H Bernhart, Ivo L Hofacker, and Peter F Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics (Oxford, England)*, 22(5):614–5, March 2006.
- [92] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10):1278–84, October 2007.
- [93] Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic acids research*, 29(5):1034–46, March 2001.
- [94] Stephan H Bernhart, Ullrike Mückstein, and Ivo L Hofacker. RNA Accessibility in cubic time. *Algorithms for molecular biology : AMB*, 6(1):3, January 2011.
- [95] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics (Oxford, England)*, 22(10):1177–82, May 2006.
- [96] Andreas R Gruber, Ronny Lorenz, Stephan H Bernhart, Richard Neuböck, and Ivo L Hofacker. The Vienna RNA websuite. *Nucleic acids research*, 36(Web Server issue):W70–4, July 2008.

- [97] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–8, April 1988.
- [98] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, October 1990.
- [99] Wolfgang Gerlach and Robert Giegerich. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. *Bioinformatics (Oxford, England)*, 22(6):762–4, March 2006.
- [100] Mirela Andronescu, Zhi Chuan Zhang, and Anne Condon. Secondary structure prediction of interacting RNA molecules. *Journal of molecular biology*, 345(5):987–1001, February 2005.
- [101] Hakim Tafer and Ivo L Hofacker. RNAPlex: a fast tool for RNA-RNA interaction search. *Bioinformatics (Oxford, England)*, 24(22):2657–63, November 2008.
- [102] Can Alkan, Emre Karakoç, Joseph H Nadeau, S Cenk Sahinalp, and Kaizhong Zhang. RNA-RNA interaction prediction and antisense RNA target search. *Journal of computational biology : a journal of computational molecular cell biology*, 13(2):267–82, March 2006.
- [103] Marc Rehmsmeier, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA (New York, N.Y.)*, 10(10):1507–17, October 2004.
- [104] TF Smith and MS Waterman. Identification of common molecular subsequences. *J. Mol. Biol*, 147:195–197, 1981.
- [105] Hakim Tafer, Fabian Amman, Florian Eggenhofer, Peter F. Stadler, and Ivo L. Hofacker. Fast Accessibility-Based Prediction of RNA-RNA Interactions. *Bioinformatics*, page 8, 2011.
- [106] The Gene and Ontology Consortium. The Gene Ontology project in 2008. *Nucleic acids research*, 36(Database issue):D440–4, January 2008.
- [107] John Day-Richter, Midori a Harris, Melissa Haendel, and Suzanna Lewis. OBO-Edit—an ontology editor for biologists. *Bioinformatics (Oxford, England)*, 23(16):2198–200, August 2007.
- [108] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, L.J. Goldberg, Karen Eilbeck, Amelia Ireland, C.J. Mungall, and Others. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

- [109] Michael Ashburner, C.A. Ball, J.A. Blake, David Botstein, Heather Butler, J.M. Cherry, A.P. Davis, Kara Dolinski, S.S. Dwight, J.T. Eppig, and Others. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.
- [110] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
- [111] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics (Oxford, England)*, 22(13):1600–7, July 2006.
- [112] Shuo Chen, Aixia Zhang, Lawrence B Blyn, Gisela Storz, Shuo Chen, Aixia Zhang, Lawrence B Blyn, and Gisela Storz. MicC , a Second Small-RNA Regulator of Omp Protein Expression in Escherichia coli MicC , a Second Small-RNA Regulator of Omp Protein Expression in Escherichia coli. *Society*, 186(20), 2004.
- [113] Liron Argaman and S Uppsala. The Small RNA IstR Inhibits Synthesis of an SOS-Induced Toxic Peptide. *Current*, 14:2271–2276, 2004.
- [114] Brian Tjaden. TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic acids research*, 36(Web Server issue):W109–13, July 2008.
- [115] Pierre Mandin, Francis Repoila, Massimo Vergassola, Thomas Geissmann, and Pascale Cossart. Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets. *Nucleic acids research*, 35(3):962–74, January 2007.
- [116] Ulrike Mueckstein, Hakim Tafer, Stephan H Bernhart, Maribel Hernandez-Rosales, Jorg Vogel, Peter F Stadler, and Ivo L Hofacker. Translational Control by RNA-RNA Interaction : Improved Computation of RNA-RNA Binding Thermodynamics. *Communications in Computer and Information Science*, 13:114–127, 2008.
- [117] IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics (Oxford, England)*, 24(24):2849–56, December 2008.
- [118] Hamidreza Chitsaz and Rolf Backofen. biRNA: Fast RNA-RNA binding sites prediction. *Algorithms in Bioinformatics*, pages 25–36, 2009.
- [119] Sandra C Viegas, Inês J Silva, Margarida Saramago, Susana Domingues, and Cecília M Arraiano. Regulation of the small regulatory RNA MicA by ribonuclease III: a target-dependent pathway. *Nucleic acids research*, 39(7):2918–30, April 2011.

- [120] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic acids research*, 32(Database issue):D277–80, January 2004.

Lebenslauf

Persönliche Daten

Eggenhofer Florian

Geb. am 17. Oktober 1983 in Tulln an der Donau, Österreich

Österreichischer Staatsbürger, ledig

evangelisch, A.B.

Mutter: Dr. Eva Eggenhofer; Vater: Ing. Peter Lehner

Schulbildung

1990-1994 Volksschule Stockerau Ost

1994-1998 Bundesgymnasium Stockerau Unterstufe

1999-2004 Private HTL für Lebensmitteltechnologie, Hollabrunn

06/2004 Matura

Hochschulbildung

10/2004 Beginn des Studiums Molekulare Biologie, Universität Wien

02/2010-11/2011 Diplomarbeit am Institut für Theoretische Chemie, Universität Wien, Titel: "RNApredator: A web-based tool to predict small RNA targets"

Lehrtätigkeit

03/2009-07/2009 Tutor für UE Molekulare Biologie IB, Universität Wien

10/2009-02/2010 Tutor für UE Molekulare Biologie IB, Universität Wien

03/2010-07/2010 Tutor für UE Molekulare Biologie IB, Universität Wien

Berufliche Tätigkeiten

02/2009–Jetzt Systemadministrator am Institut für theoretische Chemie, Wien

Sonstiges

10/2003–09/2004 Präsenzdienst im Österreichischen Bundesheer

Wien, 07. November 2011