Computational analyses of post-transcriptional regulatory mechanisms

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat
der Technischen Fakultät
der Albert-Ludwigs-Universität Freiburg

August 22, 2014

von
M.Sc. Bioinformatikerin
Sita Johanna Saunders

geb. Lange
Rawene, Neuseeland 1983

Dekan:
Prof. Dr. Yiannos Manoli

Gutachter:

1. Prof. Dr. Rolf Backofen
   Bioinformatics
   Department of Computer Science
   Albert-Ludwigs-University Freiburg
2. Prof. Dr. Wolfgang R. Hess
   Genetics & Experimental Bioinformatics
   Institute of Biology II
   Albert-Ludwigs-University Freiburg

Acknowledgements

"Life is like riding a bicycle, in order to keep your balance, you must keep moving"—Albert Einstein.

Here is to all my memories of bicycle riding in Freiburg and to all who kept me moving in my continuous balancing act.

High quality research is never a one-man—not even a one-woman—job. Having established this as a fact of life, I would like to express my gratitude for the many and fruitful collaborations I enjoyed throughout my Ph.D. years; without your collaborative effort, the majority of this work would not have been possible. Of course, Prof. Dr. Rolf Backofen, my Ph.D. supervisor, was a constant support along the way. I also appreciate that Prof. Dr. Wolfgang Hess, Prof. Dr. Christoph Scholl and Prof. Dr. Matthias Teschner agreed to be a part of my examination committee to discern the quality of my work.

In particular, I would thank...
...Fabrizio for generously sharing your fountain of ideas,
...Daniel for making the boat we found ourselves in more buoyant,
...Micha, Kyanoush, and Peter for being cool and capable students,
...Omer/Omar/Amri for working so much—as if you were split into three,
...Rhodri, Christina, Martin and Fabrizio for proofreading parts of this thesis,
...Robert, Christina, and Daniel for making the group livelier and more adhesive,
...my family and friends for not asking too frequently when I will finish my Ph.D.,
...all past and present group members for the fun and supportive group atmosphere,
...Martin for making us all take breaks once in a while and not forgetting to have fun,
...Monika for being the heart of the group and your generous help in all administration,
...Stefan for support when it came to my dreaded area of non-expertise: system configurations,
...Rhodri for being my best critical sounding board, your superb editing skills, and not least your love and support throughout these entirely stressful years.

Finally, I thank my years in research for teaching me one fundamental concept: *"The only true wisdom is in knowing you know nothing."*—Socrates

Contents

Summary

This is a dissertation about computational and statistical analyses of mechanisms in post-transcriptional gene-expression regulation. Gene expression is the process by which the genetic information, stored in a segment of the genome, is used to synthesise a functional gene product; it involves complex regulation at multiple levels. Whereas regulatory control at the DNA level usually involves an on/off mechanism, regulation at the RNA level is more varied and allows for fast and flexible adaptation to changing environmental pressures. Post-transcriptional regulation of RNA generally requires interactions between the RNA and *trans*-encoded factors, such as other RNAs or RNA-binding proteins. Although interactions frequently occur by chance, they are of little consequence unless the affinity between the RNA and its interacting partner is sufficient to facilitate a binding strong enough for the regulatory process to proceed. Two main properties of RNA affect binding affinities: the nucleotide sequence and its structure. While a *trans* factor can form interactions with specific nucleotides or sequence of nucleotides, the RNA structure can either enable better access to—or block—the active binding site. An active binding site is called a regulatory recognition element.

Although some of the presented work is applicable to a broader analysis of post-transcriptional, regulatory mechanisms, most work is applied to two popular, regulatory systems in which small RNAs interact with associated proteins to target nucleic acids and suppress their expression. In the prokaryotic CRISPR-Cas adaptive immune system, the crRNA is processed from CRISPR RNA and subsequently guides an associated complex of Cas proteins to target foreign genetic material for immediate degradation. The other system involves the microRNA in eukaryotes, which (in its mature form) is integrated into an Argonaute protein where it binds to a target RNA and causes either its degradation or storage for later use.

Sequence and structure conservation throughout evolution is a powerful indicator of non-coding RNA function: it is frequently used to classify RNAs into functional groups. CRISPR-Cas systems are extremely versatile in their mechanistic processes. Current classification of subtypes is focussed on associated sets of Cas proteins, whereas the evolution of CRISPR RNA is disregarded. In this work, we supplement Cas-protein–based subtype classifications with a comprehensive analysis of patterns in CRISPR-sequence and -structure conservation. We developed a web server that automatically assigns newly-sequenced CRISPRs to predetermined sequence families and structure motifs, and visualises CRISPR conservation in a single glance. This is the first resource to explore CRISPR-Cas systems based on CRISPR evolution.

To perform its regulatory function, non-coding RNA must first be transcribed from the DNA and processed into its mature form. An accepted method for the analysis of RNA transcripts is to apply a sequencing protocol to purified RNA called `RNA-seq`, which we used to characterise CRISPR RNA expression. Mapping `RNA-seq` reads to CRISPR loci displayed high abundances of mature crRNAs in the cyanobacterium *Synechocystis* sp. PCC6803. Furthermore, in-depth analysis of `RNA-seq` reads determined exact processing sites and indicated that highly structured crRNAs could be degraded more quickly.

Research into regulatory recognition elements frequently involved the prediction of local RNA structure. To study this aspect, we compared the performance of available structure-prediction algorithms to detect local structure in messenger RNAs on two large, independent datasets: assessing both the prediction of exact base pairs and general single-strandedness of sequence regions. We determined optimal settings for locality parameters of existing tools and developed an approach that eliminated prediction bias that arose at artificial window termini. With this work, we give the first comprehensive guide on how to predict local structures and fold long RNAs. In an application to CRISPR RNA, we developed a tool to identify the regulatory structure motif that is folded with the highest predicted stability across multiple instances. We also confirmed that the context surrounding the regulatory structure motif in CRISPR RNA affects its stability, which subsequently influences the binding affinity and cleavage activity of the respective Cas endoribonuclease.

Successful computational prediction of regulatory recognition elements has been an extremely elusive task to date. To extend on the idea that context is influential, we performed a statistical analysis of independent regions surrounding microRNA recognition elements. We identified clear signals of increased structural accessibility and nucleotide frequencies downstream of recognition elements in plants; similar signals were reflected in human and firefly data. Furthermore, we developed a machine-learning framework based on a graph-kernel that is able to capture and learn complex sequence and structure features from any class of regulatory recognition elements. The superior performance of our approach in detecting RNA-binding-protein recognition sites was established. Although its application to microRNA interaction data is not yet complete, initial results were promising.

The final aspect covered in this thesis is the design of artificial RNA for the targeted control of arbitrary gene expression on the post-transcriptional level. Here, we discovered favourable characteristics of artificial microRNAs in a model plant (*Arabidopsis thaliana*). We considerably improved the specificity of designed microRNAs by filtering results from a state-of-the art design platform using hybridisation characteristics of the interaction. In addition, we designed experiments to show that the context surrounding a target site of an artificial microRNA can enhance or inhibit its repression efficiency.

Zusammenfassung

Diese Dissertation befasst es sich mit der computergestützten und statistischen Analyse von Mechanismen, die in der post-transkriptionellen Regulation der Genexpression aktiv sind. Genexpression ist ein Prozess worin genetische Information, gespeichert auf einem Genomsegment, als Anleitung benutzt wird, um ein Genprodukt herzustellen. Dabei wird dieser Prozess auf mehreren Ebenen in komplexer Weise reguliert. Während die Regulation auf DNA Ebene generell einem An/Aus-Mechanismus folgt, ist die Regulation auf RNA Ebene dagegen vielfältiger und erlaubt eine schnelle und flexible Anpassung an sich stets verändernde Einflüsse der Umgebung. Die post-transkriptionelle Regulation von RNA benötigt für gewöhnlich eine Interaktion zwischen der RNA und einem trans*-enkodierten Faktor (Transfaktor) wie zum Beispiel einer weiteren RNA oder einem RNA-bindenden Protein. Obwohl Interaktionen häufig zufällig stattfinden, bleiben die meisten ohne Konsequenz, da sie nicht lang genug anhalten. Wenn die Affinität zwischen der RNA und ihrem Interaktionspartner stark genug ist, wird eine stabile Bindung ermöglicht und eine Regulation kann stattfinden. Zwei Eigenschaften der RNA beeinflussen die Bindeaffinität: ihre Nukleotidsequenz und ihre Struktur. Während ein Transfaktor spezifisch an Nukleotiden, oder an eine Sequenz von Nukleotiden, binden kann, wird die aktive Bindestelle von der RNA-Struktur entweder blockiert, oder sie ermöglicht einen besseren Zugang. Eine aktive Bindestelle wird ein regulatorisches Erkennungselement genannt.*

Obgleich ein Teil dieser Arbeit allgemein für die Analyse von post-transkriptionellen, regulatorischen Mechanismen anwendbar ist, befasst sie sich großteilig mit zwei spezifischen regulatorischen Systemen. In beiden Fällen agieren kleine RNAs zusammen mit assoziierten Proteinen und behindern die Expression von Genen indem sie gezielt regulatorische Erkennungselemente binden. Im prokaryotischen, adaptiven CRISPR-Cas Immunsystem werden mehrere kleine crRNAs aus einer langen CRISPR RNA prozessiert. Diese crRNAs agieren zusammen mit assoziierten Cas-Proteinen, um fremdes, angreifendes, genetisches Material zu zerstören. Das zweite System umfasst microRNAs in Eukaryoten. Eine microRNA (in ihre ausgereifte Form) wird in ein Argonaut-Protein integriert, um eine Ziel-RNA zu binden. Dies verursacht ihre Degradation oder Speicherung für einen späteren Zeitpunkt.

Die Konserviertheit von Sequenz und Struktur, trotz evolutionsbedingte Diversität, ist ein wichtiges Merkmal funktionsfähiger nicht-kodierender RNA: sie wird oft für die Klassifizierung

von funktionalen Gruppen in RNA verwendet. CRISPR-Cas-Systeme sind äußerst vielfältig in ihren zugrundeliegenden Mechanismen. Die gängige Klassifizierung von Subtypen ist auf die assoziierten Cas-Proteine eines Systems fixiert, wobei die Evolution von CRISPR RNA nicht berücksichtigt wird. In dieser Arbeit ergänzen wir die vorhandene Cas-Protein-basierte Subtypklassifizierung mit einer umfassenden Auswertung der Konservierung von Sequenz und Struktur in CRISPR RNA. Wir entwickelten einen Webserver für die automatische Zuteilung von neu-sequenzierten CRISPR RNAs zu unseren vorbestimmen Sequenzfamilien und Strukturklassen, und wir veranschaulichten die Ergebnisse, so dass die CRISPR-Konservierung auf einen Blick erfasst werden kann. Unser Webserver bietet damit den ersten Service an, mit dem CRISPR-Cas Systeme anhand ihrer CRISPR-Evolution untersucht werden können.

Eine regulatorische Funktion kann nur erfolgen, wenn die zuständige nicht-kodierende RNA von der DNA zuerst abgelesen (transkribiert) und anschließend zu ihrer funktionalen Form prozessiert wird. Ein allgemein anerkannter Ansatz RNA-Transkripte zu analysieren, ist die Verwendung eines Sequenzierungsprotokolls für aufgereinigte RNA (`RNA-seq` genannt). Diesen Ansatz haben wir auch für die Bestimmung von CRISPR-RNA-Transkripten verwendet. Die Zuordnung von sequenzierten "Reads" zu CRISPR-Genen in Synechocystis *sp. PCC6803 hat eine sehr hohe Expression von prozessierten crRNAs offenbart. Desweiteren entdeckten wir mittels einer vertieften Auswertung der `RNA-seq`-Daten exakte Prozessierungsstellen und einen Hinweis darauf, dass stark-strukturierte crRNAs schneller degradiert werden.*

Die Forschung von regulatorischen Erkennungselementen benötigt häufig eine Vorhersage von RNA Struktur. Wir entschlüsselten diesen Aspekt, indem wir das Leistungspotenzial von verfügbaren Algorithmen für die Vorhersage lokaler Struktur in "messenger RNAs" auf zwei großen, unabhängigen Datensätzen verglichen. Die Qualität der Vorhersagen wurde sowohl für die exakte Basenpaarung als auch für die allgemeine Zugänglichkeit der Nukleotide ermittelt. Wir haben die optimalen Einstellungen der Lokalitätsparameter bestimmt und einen Ansatz entwickelt, um inkorrekte Vorhersagen an artifiziellen Fensterenden zu eliminieren. Mit dieser Arbeit liefern wir eine erste umfassende Richtlinie wie lokale Struktur vorhergesagt werden kann und wie die Faltung von langen RNAs am besten funktioniert. Ermitteltes Wissen wurde anschließend für die Vorhersage von lokalen Strukturmotiven in CRISPR RNA angewendet. Hierbei entwickelten wir eine Methode, um die stabilste Struktur von mehrfach auftretenden Strukturmotiven innerhalb eines CRISPR-Transkriptes zu bestimmen. Darüber hinaus konnten wir bestätigen, dass die Sequenz in der Umgebung eines strukturierten, regulatorischen Erkennungselements die Ausbildung der Struktur negativ beeinflussen und somit die Erkennung und Spaltung mittels der zuständigen Cas-Endoribonuklease verhindern kann.

Die akkurate computergestützte Vorhersage von regulatorischen Erkennungselementen ist ein noch ungelöstes Problem. Um auf die Beobachtung, dass die Umgebung eines regulatorischen Elements dessen Erkennung beeinflussen kann, aufzubauen, unternahmen wir eine statistische Auswertung von unabhängigen Regionen im Umfeld von microRNA-Erkennungselementen. Wir entdeckten klare Signale von erhöhter struktureller Erreichbarkeit und auffallende Nukleotidfrequenzen in der Nähe von Erkennungselementen in Pflanzen; ähnliche Signale waren

auch in Interaktionsdaten aus dem Menschen und dem Leuchtkäfer zu erkennen. Ferner entwickelten wir zusätzlich einen Ansatz, basierend auf maschinellem Lernen und einem Kernel für Graphen, welcher komplexe Sequenz- und Struktureigenschaften von regulatorischen Erkennungselementen jeglicher Art erfassen kann. Die überragende Qualität von Vorhersagen auf regulatorischen Erkennungselementen von RNA-Bindeproteinen hat sich bewährt. Obgleich die Performanz auf microRNA-Interaktionsdaten noch nicht vollständig ermittelt wurde, waren erste Ergebnisse vielversprechend.

Der letzte in dieser Dissertation betrachtete Aspekt ist die Konstruktion künstlicher RNA, um die Expression von frei-wählbaren Genen auf der post-transkriptionellen Ebene zu kontrollieren. Hier haben wir vorteilhafte Eigenschaften von künstlicher microRNA im Modellorganismus der Pflanzen Arabidopsis thaliana *entdeckt. Mit einem Filteransatz von Hybridisierungsmerkmalen der Interaktion konnten wir die Spezifizität der künstlichen microRNAs, die durch Standardverfahren konstruiert wurden, erheblich verbessern. Zudem haben wir Experimente aufgestellt, in denen wir zeigten, dass die Sequenzumgebung einer Zielregion der künstlichen microRNA ihre Hemmungseffizienz entweder erhöhen oder senken kann.*

Part I

Introduction

CHAPTER 1

Overview

The purpose of life is a life of purpose.—Robert Byrne

1.1 Motivation

Life would be frozen without the active regulation that is the driving force behind growth, constant adaptation to the environment and just everyday functionality. Every molecule in a living organism has a purpose. Through interaction with others, a molecule is guided and induced into action and so able to fulfil its purpose. A molecule on its own is like a car without a driver or a driver without a car—without purpose. The main principle behind this work is to elucidate the properties of both driver and car that enable their interaction during the act of driving: for example, the driver must be sitting in the driver's seat, be able to reach the steering wheel, the ignition, the pedals, all buttons and levers, and see the road over the dashboard and in the mirrors.

In a biological cell, genetic information is stored in the genome. The process by which the information in DNA segments (genes) is used to create a functional product is called gene expression. During gene expression DNA is first transcribed to RNA. A subset of these RNA transcripts are then translated into proteins and the remaining RNA transcripts are non-coding because they do not encode for a protein but act as regulators or catalysts themselves. All macromolecules, DNA, RNA, and proteins, in combination with various metabolites, work together competitively and collaboratively to form a vast network of dependencies. Thus, expression levels of each type of macromolecule are finely regulated with astounding complexity and dexterity. Gene-expression regulation is initiated by the interaction of two or more partners in either a lock-and-key or an induced-fit interaction: are the seat, the headrest and the mirrors compatible with the driver's physical build (lock-and-key), or does the driver first have to adjust these parts before starting to drive (induced-fit)? In addition, there are a multitude of both cars and drivers: a car can only be driven if both are in the same location and the driver has the corresponding key.

This dissertation introduces computational methodologies and analyses applied to multiple aspects of post-transcriptional gene regulation with a focus on sequence and structure properties of RNA that affect regulatory interactions. Due to the enormity of the field, one could fill thousands of dissertations with meaningful advances, but still not have elucidated all post-transcriptional regulatory processes. The herein presented computational methodologies and analyses can be viewed as building blocks or guidelines for solving similar questions in the future. In addition it provides noteworthy—largely published in peer-reviewed journals—advances to selected biological applications.

Many people question what computer science has to do with the study of molecular biology. Modern research relies heavily on the application of computers to solve complex problems, predict probable outcomes that can later be tested in a laboratory, or analyse vast data sets. Computers have enabled research to move away from single examples to look at whole systems. In the past two decades, research of molecular biology has made monumental progress with the development of automated processes and high-throughput, experimental techniques that generate plentiful data every day. For example, take the progress made in the unravelling of genetic code and its products: sequencing the human genome took about 20 years with the last human chromosome completed in 2006 [110]. Today, the time and cost of high-throughput or genome sequencing has dropped so dramatically that applications to personalised medicine are being considered [221]. These second- and third-generation sequencing technologies also allow the comparison of tissues at various time points or the detection of potentially disease-linked genetic signatures [169]. Over 40 million protein sequences are available from 41,263 species (status on 15.07.2014 from the National Center for Biotechnology Information Reference Sequence Database, `http://www.ncbi.nlm.nih.gov/refseq/`) with exponentially growing numbers. In one experimental run, advanced mass spectrometry techniques detect hundreds to thousands of proteins (and metabolites) present in a sample [247]. Analysis, annotation and storage of this large volume of information would not be possible without (computer) scientists that develop software or specialised algorithms to perform complex computational or statistical analyses.

1.2 General objectives

Within the genome, a gene encodes a functional gene product—either a protein or a non-coding RNA (ncRNA). The process of generating the functional gene product is called gene expression. Regulation of gene expression occurs first on the DNA level and subsequently on the RNA level. Although the regulation of gene expression has been extensively explored since the discovery of the genetic code, especially on the DNA level [177, 239], this process is so complex that our understanding is, metaphorically speaking, still in the stone ages. The focus of this dissertation is on *computational* analyses of mechanisms in post-transcriptional regulation, which is the control of gene expression at the RNA level between transcription and translation. Post-transcriptional control generally involves factors encoded in *trans* and/or in *cis* on the RNA transcript. In this work, we analyse common processes in which *trans* factors, e.g. an RNA-binding proteins (RBP) or ncRNAs, bind specifically to corresponding regulatory recognition elements with RNA transcripts to facilitate their regulatory control. Particular focus was put on two RNA-based regulatory systems that have shown great potential for applications in biotechnology. First, we investigated the adaptive immune response in prokaryotes that is provided by the CRISPR-Cas system, in which a small RNA (∼45–70 nt) performs a central regulatory role in defending the organism against foreign genetic material. In this work, we analysed conservation, expression and processing of the CRISPR RNA in CRISPR-Cas systems. The second regulatory system is called RNA interference. Again, a small (∼20 nt) RNA (called miRNA) is the central factor in regulating the expression of an endogenous target gene by binding to its mRNA. Here we explored properties of RNA sequence and structure that determine miRNA-based regulatory function.

In addition to several individual data analysis tasks that provide biological insights into properties of regulatory RNA, we advanced the state of the art of bioinformatics approaches in the following areas:

- On a collection of all available CRISPRs from public databases, we performed a comprehensive analysis of CRISPR conservation, and combined with an easy-to-use web server, we provide the first computational tool for comparing systems based on the central RNA element.

- We compared RNA structure prediction approaches on large, curated datasets to demonstrate how to predict local RNA structure in mRNAs with the highest prediction accuracy; we especially gave insights into the effects of parameters used for local structure prediction.

- We developed an efficient machine-learning framework to flexibly capture binding preferences of RNA regulatory recognition elements. In this work, we applied the framework to modelling miRNA recognition elements.

Despite its computational nature, this thesis also has a strong biological focus. The general aim was to develop and apply computational approaches to support and complement biological research on RNA regulatory mechanisms. Numerous collaborations with wet-lab experimental groups showcase the applicability of presented work to solving biological problems. In fact, 7 out of the 11 publications based on this work were produced in close collaboration with at least four separate wet-lab experimental groups (a list of my publications based on this thesis can be found after the Appendix and before the general bibliography).

1.3 Thesis guide

The collective work presented in this dissertation touches on many aspects of post-transcriptional gene regulation; it is not a simple one-topic, one-answer piece of research. Therefore, additional effort has gone into structuring the dissertation in such a way that it is not necessary to read each and every piece of work, and one can focus on sections of interest. For a general overview, it is possible to just read the introductions of each part and all conclusions—both chapter-specific conclusions and the final remarks section.

To enable a better structure of presented work, the dissertation has been divided into chapters and parts. The individual chapters describe detailed methods and applications that solve a specific problem. To put the chapters into the context of the general topic of this thesis, they have been divided into parts:

- **Part I**: aside from the current chapter, this part includes biological and computational facts, definitions and approaches that help the reader to understand presented work.
- **Part II**: exploits evolutionary conservation of sequence and structure in prokaryotic CRISPR-Cas immune systems to identify important regulatory motifs.
- **Part III**: deals with the expression of a non-coding RNA (the CRISPR RNA) and how it is processed into its mature form.
- **Part IV**: answers the question of identifying local structure in long RNA sequences, which is required for characterising many regulatory functions.
- **Part V**: focusses on characterising regulatory binding sites within such long RNAs (e.g. mRNAs).
- **Part VI**: explores the design of artificial RNA that can be used for targeted post-transcriptional regulation of specific genes.
- **Part VII**: concludes the entire thesis, explains general limitations and offers ideas about future work.

In the appendix, the reader can first find a glossary of commonly used terminology, a declaration of the authenticity of this work, a point-by-point statement of contributions and additional material for each of the parts. Much of the work presented in this thesis was published in peer-reviewed journals (or is currently under review). However, some of Part V and all of Part VI has not been made publicly available previous to this dissertation. My own publications are separated from the remaining references in the bibliography and in the text, the difference in citation style can be used to differentiate own publications from others: e.g., [P3] references a publication of which I am an author and [239] is the style used for other references.

1.4 Statement of contribution

As established in my acknowledgements, high-quality research in the modern era is never achieved by just one person. Throughout my Ph.D. years, I worked with many people from both within the Freiburg Bioinformatics Group and various other external individuals and groups.

I have made every effort to reduce the amount of external contributions presented. When extracting external contribution would be detrimental to understanding the presented research, however, it was retained. Detailed statements of contribution and sources of presented work, when taken from my own publications are provided in Appendix C: from these statements, my own contribution is clarified. I would like to note that I contributed substantially or solely to the writing of all parts in this dissertation that were taken partly from publications of which I am an author.

Finally, I come to the use of *we* as the chosen pronoun throughout this work. Although *I* was significantly involved in the development of proposed methods and all presented analyses (if not otherwise referenced), in acknowledgement of my collaborators, *we* is appropriate. For consistency reasons, we is used even if the work was done solely by myself.

CHAPTER 2

Biological and computational background

This chapter provides an overview of biological and computational aspects of RNA-based post-transcriptional gene regulation. First, the major players in post-transcriptional gene regulation are introduced. In addition, popular regulatory mechanisms, which are highly relevant to medicine and biotechnology, are investigated throughout this dissertation and are described in more detail herein. Second, a broad insight into state-of-the-art computational approaches that are key to the research of regulatory RNA are presented: approaches to determine aspects of RNA structure, conservation, regulatory RNA and interactions with RNA are covered.

2.1 Post-transcriptional regulation of gene expression

Post-transcriptional regulation is the control of gene expression at the level of RNA, which occurs after transcription of DNA into RNA and before translation of RNA into protein[1] (Figure 2.1). This RNA-based regulation of gene expression frequently involves the binding of *trans*-acting–regulatory factors (*trans factors*) to a longer RNA, frequently messenger RNAs (mRNAs). There are three major factors to consider: (1) the *cis*-encoded regulatory recognition element (RRE) on the RNA being regulated, (2) characteristics of the *trans* factor that facilitates binding, and (3) stoichiometric effects of differential expression levels. Computational predictions are only possible for the first two factors; the last factor requires experimental measurements of expression levels. The ultimate goal of post-transcriptional regulation is to control (with RNA) the processing, transport, localisation, number and variants of gene products in a cell—of both protein-coding and non-coding genes. Key molecules involved in post-transcriptional regulation are mRNA, regulatory non-coding RNA (ncRNA), and RNA-binding proteins (RBPs).

[1] The control of gene expression at the level of proteins is called post-translational regulation.

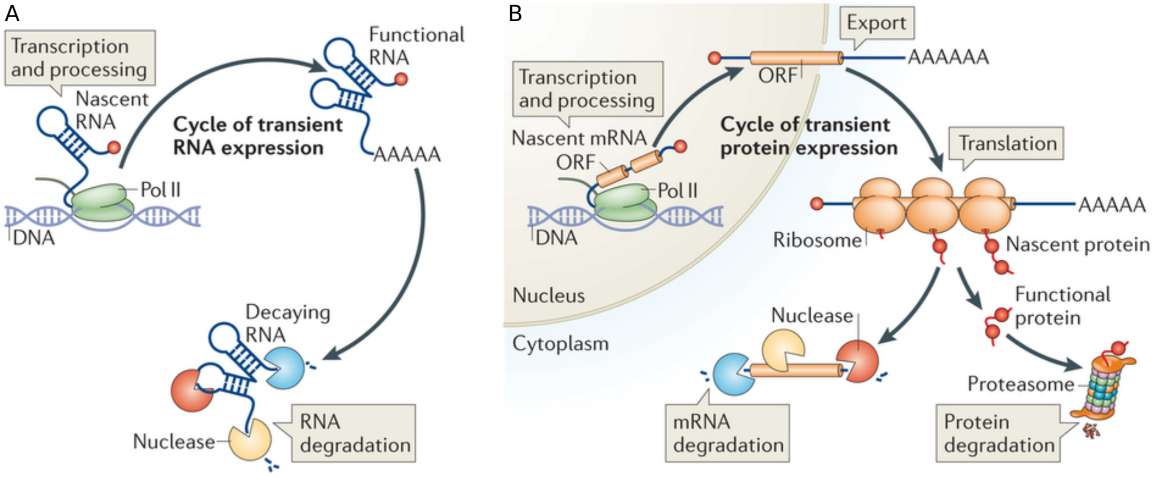


Figure 2.1. RNA expression cycle. RNA is expressed as (A) non-coding RNA (ncRNA) or (B) coding RNA (mRNA). The life cycle of expression and degradation is illustrated for a general eukaryotic cell; this life cycle is controlled by a hyper-complex network of regulatory process that act on the DNA, RNA and protein level for either large-scale or fine-tuning effects—for fast and transient reactions to the environment or for more permanent changes. Illustration adapted with permission from Macmillan Publishers Ltd: *Nature Reviews Molecular Cell Biology* [101], copyright 2013 (license no. 3363740908111).

2.1.1 Messenger RNAs

The mRNA is the carrier of the message of genetic code from the protein-coding gene on the genome to the ribosome where the mRNA is translated into a sequence of amino acids (polypeptide). The existence of an mRNA in a cell begins with transcription and ends in degradation (Figure 2.1). During its life cycle, an mRNA molecule may be processed, edited, and transported prior to translation. There are several characteristic differences between prokaryotic and eukaryotic mRNAs; major differences are as follows: (1) eukaryotic mRNAs usually require extensive processing and transport from the nucleus to the site of translation, whereas, prokaryotic mRNAs are mostly translated while still being transcribed as there is no nucleus in a prokaryotic cell; (2) the lifetime of a prokaryotic mRNA is much shorter; and (3) prokayrotic mRNA can encode several genes at a time (called polycistronic mRNA), whereas eukaryotes only ever encode one gene (monocistronic). After all processing steps, the general form (for all purposes in this thesis) of a mature mRNA is monocistronic with a 5' (left terminus in Figure 2.2) cap and a poly(A) tail, which is a sequence of consecutive adenosine (A) nucleotides attached to the 3' (right terminus in Figure 2.2) end, to protect it from degradation. The coding sequence (CDS) is initiated by a start codon and ends with a stop codon and is a multiple of three nucleotides where each triplet encodes a single amino acid. The CDS is flanked by untranslated regions (UTRs) at the 5' and the 3' ends, called the 5'UTR and 3'UTR, respectively. The UTRs, especially the 3'UTR, contain several *cis*-regulatory elements (or RREs) with conserved structure motifs (Figure 2.2). A database of *cis*-regulatory elements is provided by Jacobs and colleagues [149]. Other RREs are small binding sites (∼4–25 nt) where the interaction with specific ncRNAs or RBPs occur (Figure 2.2). Although regulatory motifs occur predominantly in the UTRs some are also found in the CDS [250].

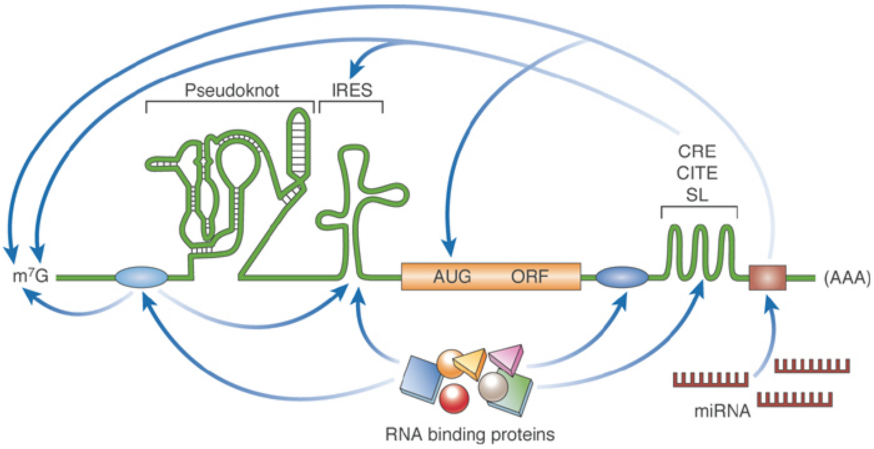


Figure 2.2. Illustration of regulatory recognition elements that control mRNAs. Gene expression is regulated on many levels. Here, the two types of regulatory elements are illustrated by examples that control mRNA translation: (1) *cis*-regulatory elements, which are either structured or unstructured motifs on the same mRNA that either act as binding sites or act directly to control its own translation and (2) *trans*-regulatory elements, which are ncRNA, RNA-binding proteins, or other molecules that bind specifically—or unspecifically—to regulate mRNA translation, degradation, or localisation. IRES (internal ribosome entry site) elements are found in the 5'UTR and promote cap-independent translation. Complex pseudoknot structures, located anywhere within the mRNA can affect translation initiation, frame shifting or termination. Among many others, structured *cis*-regulatory elements, generally in untranslated regions (UTRs), such as CREs (*cis*-acting replication elements), CITEs (cap-independent translational enhancer) and SLs (stem-loops) often influence the mRNA via long-range RNA interactions. MicroRNAs (miRNAs) bind to their respective target sites and via RNA-RNA interaction and together with associated proteins, they inhibit translation or degrade target mRNA. Several regulatory mechanisms target the mRNA cap (7-methyl-guanosine—m^7G), the *AUG* translation initiation codon and the poly(A) tail as these are common features of mRNAs. The blue ovals and brown rectangles are RBP and microRNA binding sites, respectively. Details were taken from [264]; The illustration was adapted with permission from John Wiley and Sons: *EMBO REPORTS* [264] copyright 2009 European Molecular Biology Organization (license no. 3363740305420).

2.1.2 Non-coding RNAs

Only about 2 % of the human genome is comprised of protein-coding genes [53]; the remaining 98 % was initially considered to be "junk DNA". Although we now know that the genome is pervasively (>90 %) transcribed into RNA [54], we still do not understand the function of most transcripts. Although there is some evidence that many transcripts are not functional [319].

Functional RNA transcripts that do not contain protein-coding information are termed non-coding RNA (ncRNAs); see [104, 209] for an overview of ncRNA function. Many important genes fall into the non-coding category, for example, transfer RNA (tRNA) carries its respective amino acid to the site of translation [4, 40]; ribosomal RNA (rRNA) makes up the translation machinery, the ribosomes [51, 351]; microRNAs (miRNAs) and small interfering RNAs (siRNA) that regulate gene expression by inhibition or degradation of mRNAs [6]; clustered regularly interspaced short palindromic repeats (CRISPR) that guide an adaptive defence mechanism in prokaryotes [2, 307, 337]. Most of the presented research in this dissertation deals with regulatory mechanisms involving miRNAs and CRISPRs, which are discussed further in Section 2.2.

The annotation of ncRNAs remains incomplete. Bioinformatic approaches detect hundreds of previously unknown, structured putative ncRNAs that are conserved across many species, which still require further characterisation [335]. Most known ncRNA species form evolutionarily conserved (global) structures. In contrast to protein-coding RNA, the structure of ncRNA is generally more conserved than the nucleotide sequence. Sequence-and-structure conservation of ncRNAs form the basis of detection algorithms (*c.f.* Section 2.6). However, not all ncRNAs form conserved global structures and these are more difficult to detect.

2.1.3 RNA-binding proteins

Most post-transcriptional regulatory processes, such as splicing, polyadenylation, processing, stabilisation, localisation are controlled by a class of proteins that contain an RNA-binding domain, namely RBPs (RNA-binding proteins) [106]. Recent efforts were made to identify the magnitude of the human RBPome (i.e., all RBPs encoded in a genome) where more than 800 RBPs were identified in humans [12, 38]. In yeast it was established that at least 72 % of protein-coding genes were bound by RBPs and that RBP-target sites are highly conserved [91]. These results point to a complex and vast network of RBP-regulated processes, however, relatively few RBPs are well characterised. In fact, over 300 of the RBPs identified in the first two studies [12, 38] were previously unknown, let alone fully characterised.

RBPs often display highly selective binding to their target RRE sites [199, 297]. Targets are mostly mRNAs or ncRNAs; almost all ncRNAs function together with RBPs as a ribonucleoprotein complex. Although all RBPs bind RNA, binding strengths vary such that some interactions are transient, whereas others last the entire lifetime of the RNA. Determining binding affinities and target sites of RBPs is central to the research of post-transcriptional regulation because of their ubiquitous involvement. RBPs display different specificities for nucleotide sequences, for example an alternative splicing factor, TIA-1, binds to U-rich regions [89, 168]; SFRS1 binds to a GA-rich pattern [302]; PTB bins CU-rich sites [242]; and RBPs involved in the stabilisation and destabilisation of mRNAs bind to AU-rich regions [14, 26, 241]. Many RBPs do not only display sequence-specific affinities, but are also specific to a structural context. The most simple structural contexts are regions on the RNA that do not form intramolecular base pairs and are thus accessible for binding (unpaired regions). The prokaryotic global regulator, Crc, binds to an A-rich unpaired region [220]. In addition, experimental evidence suggests that RBPs show sequence specificity when binding to not only unpaired but also to paired regions. Lee and colleagues [185] analysed RNA sequences that bound to the C5 protein and identified a hairpin structure motif that together with the sequence was essential for C5 binding. Another example is TRBP, a human protein that binds the immunodeficiency virus type 1 TAR RNA. The use of RNA probe-shift assays showed that TRBP binds preferentially to double-stranded regions rich in guanines and cytosines [100]. In prokaryotes, an endoribonuclease involved in the CRISPR-Cas immune response, preferentially binds to one side of a small hairpin stem that contains mainly cytosines [129].

2.1.4 Transcriptomics and experimental detection

Post-transcriptional gene regulation is often analysed by looking at all RNA transcripts that are present in the cell at a given time point (called *transcriptomics*). A well-established technology for detecting expression levels of transcribed RNAs are microarrays [163,238,322]. A microarray involves a time-consuming and potentially expensive process where short oligonucleotide probes that represent every gene in the genome (often several probes per gene are required) have to be designed and produced. These oligonucleotide probes are fixed to a solid substrate where they bind to their target RNA transcripts via base-pair complementarity. Relative numbers of bound transcripts can be quantified by measuring fluorescence intensities of bound transcripts. Once a microarray has been produced, repeated analyses on that array are efficient and is thus well suited to industrial or standard routines. To complement microarrays, next-generation high-throughput sequencing techniques [221] are becoming more and more common in the application to transcriptomics. Sequencing techniques display a pronounced adaptability, well suited to the high-paced nature of research and varying model organisms [294]. RNA sequencing, often referred to as `RNA-seq`, is not only used to detect which genes are expressed at given time points, but has a broad range of notable applications (see [294] for a review). High-throughput sequencing methods and transcriptome-wide applications have been a valuable source of new data on many aspects of post-transcriptional gene regulation. These system-wide measurements have facilitated major advances in the past few decades in the fact that they have enabled computational analyses of regulatory interactions between *trans* factors and RNA transcripts (c.f. Section 2.7).

2.2 Popular, RNA-based regulatory systems

Two RNA-based regulatory systems have been of particular interest to researchers of molecular cell biology in the last decade: RNA interference (RNAi) in eukaryotes and the CRISPR-Cas prokaryotic immune system. Both systems integrate a short, guide ncRNA into a complex of associated proteins to form the regulatory effector complex that targets RNA or DNA for gene silencing or for protection against foreign genetic. This silencing mechanism allows for powerful applications in biotechnology and medicine. Although many families of short ncRNA can be used as a guide in RNAi (e.g. siRNAs [181]), mainly miRNAs are considered here. A general overview of RNAi is given in [87,128,181].

2.2.1 RNA interference with microRNAs

MicroRNAs (miRNAs) are a widespread and conserved family of ncRNAs used in RNAi to regulate gene expression at the post-transcriptional level in plants, animals and some viruses (see [6] for a recent review). The latest release of the major miRNA databank, `miRBase` (version 20, June 2013), contains nearly 25 thousand miRNA genes from more than 200 species [176], which give rise to at least 30 thousand mature miRNA products. More than half of all protein-coding genes in mammals have been identified to be evolutionarily conserved targets of miRNAs [92]; a clear indication of their pervasive regulatory impact. After their discovery in 1993 [186] and despite the many intelligent minds dedicated to miRNAs and their functions, we are only beginning to understand the diverse mechanisms of miRNA-based regulation [6]. One of the most elusive problems is the (computational) detection of miRNA target genes (Section 2.7.3). Although the effect of repression on most target genes is tiny in comparison with regulatory mechanisms prior to transcription [313], miRNAs collectively have a significant impact on nearly all cellular pathways, from cell differentiation to oncogenesis; and their malfunction is related to many serious diseases, especially cancer [44,81,200,212,293]. Hence, accurate miRNA target detection is highly sought after and some initial work was performed to this end, presented in Part V of the dissertation.

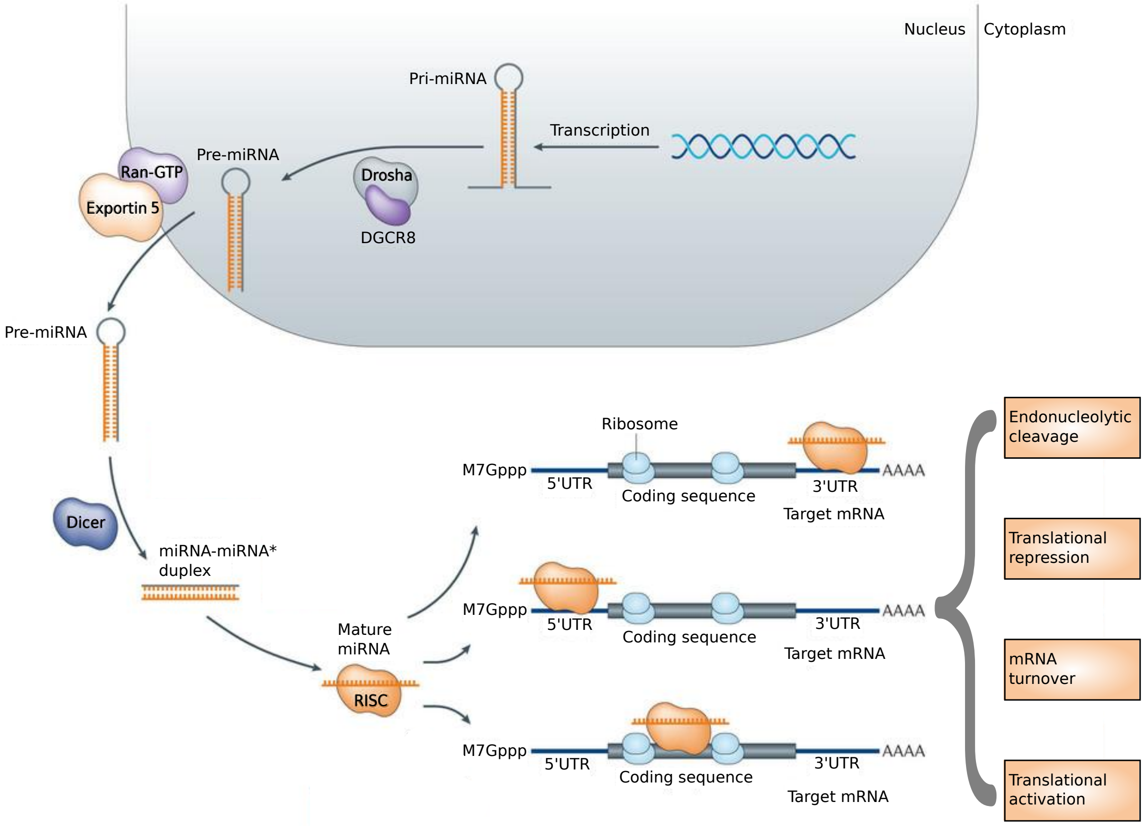


Figure 2.3. Schematic overview of miRNA biogenesis and function. The primary miRNA (pri-miRNA) is transcribed directly from the miRNA gene locus and cropped to the characteristic stem-loop structure, which is called the precursor miRNA (pre-miRNA). After the pre-miRNA is exported to the cytoplasm, it is processed further to form the miRNA-miRNA* duplex. The mature miRNA is then integrated into the Argonaute (AGO) protein and assembled into the RNA-induced silencing complex (RISC) [245]. The miRNA within the RISC guides the complex to its target and although binding to the 3'UTR has been assumed to be the predominant action, binding to the CDS is frequent and rare binding to the 5'UTR also occurs. A myriad of regulatory mechanisms exist and the main processes are summarised in the outlined boxes. The presented biogenesis is typical in animals, but differences exist, especially in plants. The main difference between animals and plants are the processing proteins and that in animals, the pre-miRNA is exported to the cytoplasm, whereas in plants the miRNA–miRNA* duplex is exported to the cytoplasm. Illustration is adapted with permission from Macmillan Publishers Ltd: *Nature Reviews Drug Discovery* [195] copyright 2013 (license no. 3363750423492).

Functional mature miRNAs are derived from a longer transcript and have a multistep biogenesis process, which can differ between species (c.f. [10]). Figure 2.3 illustrates the basic steps in the biogenesis of a mammalian miRNA. To add further complications, deep sequencing of small RNAs from a range of tissues and cell types has shown that miRNA genes produce multiple mature isoforms, known as isomirs (see [6] for a compact review on miRNA biogenesis and isoform generation). The following steps are common to all systems, only involved proteins and cell locations differ: (1) the miRNA gene is transcribed to form the primary transcript (pri-miRNA); (2) the pri-miRNA is cropped to form the characteristic stem-loop structure; this cropped transcript is termed the precursor miRNA (pre-miRNA); (3) the pre-miRNA is processed further by Dicer (or a Dicer-like protein) to form a double-stranded RNA duplex of the mature miRNA and its opponent strand miRNA*; and finally (4) an RNA-induced silencing complex (RISC) is assembled around the mature miRNA (and in some cases the miRNA*) [245]. The fully assembled RISC is then "guided" by the miRNA sequence to identify and bind a target RNA. Once bound to the target, the RISC can associate with a variety of secondary proteins to initiate or perform one of the following regulatory functions (Figure 2.3): endonucleolytic cleavage, translational repression, mRNA turnover, and sometimes even translation activation. Endonucleolytic cleavage occurs in both animals and plants, however, it is rare in animals and the predominant function in plants [6]. RISCs loaded with a miRNA (miRISCs) that do not lead to endonucleolytic cleavage either inhibit translation (translational repression) or initiate decapping and deadynylation of the target mRNA, leading to its subsequent degradation (mRNA turnover). In rare cases, miRNA targeting has been known to cause an up-regulation of the target transcript [235, 321]. Research into miRNA induced regulatory mechanisms is still ongoing, but it is clear that their mechanism of achieving a regulatory effect is flexible and diverse.

RISCs always contain a member of the Argonaute (AGO) protein family [211]. AGOs bind to any of the available small ncRNAs with no clear preference. Structural analyses have revealed that it is the conformation of the RNA-binding pocket in the AGO which determines the nucleotides that are available for RNA-RNA interaction [226, 330]. Typically, the most prominent part of the miRNA available for binding are the nucleotides between positions 2–8, which is termed the *seed*. The interaction between the seed and its target RNA is well-documented in the literature [78, 166, 226]. Beyond this seed interaction, different types of hybridisation patterns between miRNA and target exist: endonucleolytic cleavage usually requires near perfect complementarity, especially around the cleavage site between positions 10–11 of the miRNA; whereas for the other repression mechanisms, only a seed interaction can be sufficient [6, 166]. Once bound to a target, the AGO either catalyses the endonucleolytic cleavage with its own PIWI domain or it acts as a scaffold for secondary silencing factors, such as the GW-repeat containing protein GW182 [6, 66]. RISCs that do not cleave the target benefit from multiple, consecutive binding sites; such multiple binding sites lead to an increased signal and a cooperative effect on the repression activity [30].

2.2.2 The CRISPR-Cas defence mechanism in prokaryotes

Acquired immunity in prokaryotes is directed by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and their associated (Cas) proteins. This CRISPR-Cas system, present in many bacteria and most archaea, recognises and subsequently degrades exogenous genetic elements (for reviews see [2, 307, 337]). The adaptive immune response is divided into three main phases (Figure 2.4): (1) *Adaptation*, the selection of short target segments (protospacers) from foreign DNA and the incorporation of their reverse complement sequence (spacers) into the organism's active CRISPR locus between directly-repeated sequences (repeats); (2) *crRNA maturation*, expression of the CRISPR RNA and subsequent processing of the transcript into mature RNA species (crRNA); and (3) *target interference*, invader DNA [97] or RNA [125, 354] degradation at the respective protospacer, guided by the crRNA and a highly specific complex of Cas proteins such as Cmr [125, 354] or Cascade [154, 227]. The targeting complex differentiates real protospacers from other complementary sequences in many systems by a short 1–3 nt protospacer adjacent motif (PAM). In addition, the recognition of PAM motifs avoids autoimmunity, which would occur if the organism harbouring a CRISPR-Cas system would recognise its own DNA at the CRISPR locus and target it for degradation.

CRISPRs are associated with more than 50 genes [202] specific to CRISPR-Cas systems. These *cas* genes are generally encoded in single cassettes, close to the associated CRISPR locus. However, many exceptions to this rule exist, which can complicate subtype classifications. CRISPR-Cas systems are commonly classified into three types I–III and over ten subtypes, mainly by the co-occurrence of *cas* genes encoded in cassettes and generally disregarding the CRISPR RNA [123, 201, 202]. Very recently this classification has been extended further in archaea to include many variant subtypes[1] [324].

CRISPR loci are identified by their characteristic spacer-repeat architecture; the most widely used detection programs are `CRISPRFinder` [115] and `CRT` [23]. Identified CRISPRs are stored in databases, such as `CRISPI` [269] and `CRISPRdb` [114]. However, automated classification and annotation pipelines and easy-to-use characterisation of whole CRISPR-Cas systems are not yet available.

Research into this highly adaptive and diverse immune system is fairly recent; the acronym CRISPR was first proposed in 2002 [151]. Since then research into this field has made leaps and bounds with applications in biotechnology that go beyond its native function [308]. The majority of CRISPR-Cas systems are of type I or III, which are found in both bacteria and archaea; these systems are complex with about 7–10 Cas proteins and require either the large Cascade or Cmr complexes for successful defence reactions. In contrast, type II systems are light-weight with only four associated Cas proteins and are only found in bacteria. The large effector complexes required for type I and II systems are replaced in type II systems with a single gene, Cas9. Cas9, together with a *trans*-encoded tracrRNA and a guide crRNA, is

[1] This most recent classification by Vestergaard and colleagues was published after the work done in this thesis and therefore it was not integrated.

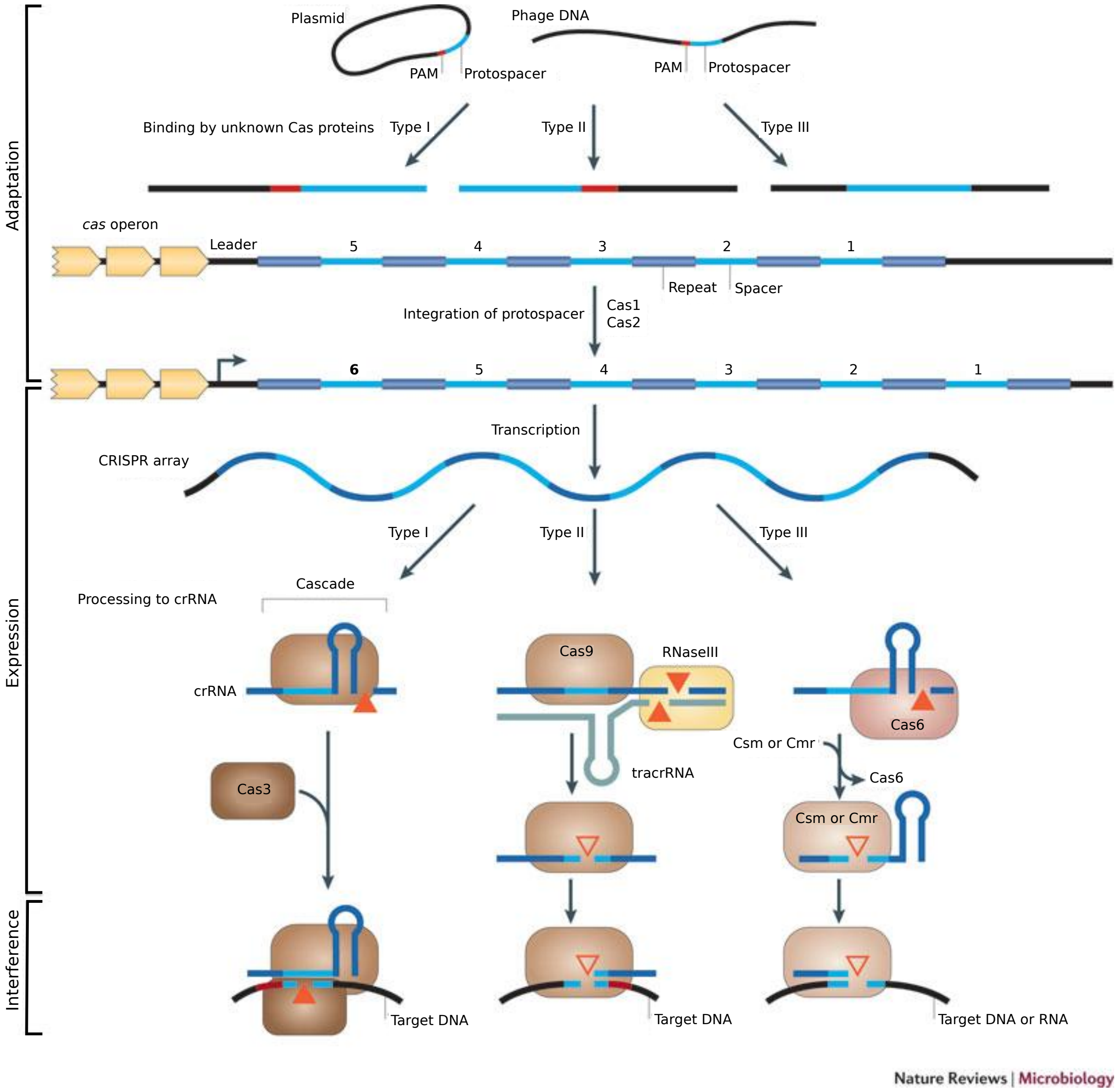


Figure 2.4. Schematic overview of the CRISPR-Cas system. In adaptation, unknown Cas proteins find suitable protospacer targets and regulate the integration of new spacers into the CRISPR locus, usually at the 5' end of the array; in most cases, the history of adaptation steps can be read from left to right, with the spacer at the 3' end being the oldest to be captured. The CRISPR array is generally expressed as a single transcript and subsequently processed into mature crRNA. The processing mechanism is highly specific and many differences exist between the systems and types: Types I and II involve a Cas6-like endoribonuclease to cleave at either a small hairpin structure motif within the repeat [32, 103, 129, 131, 227, 274, 298] or at an unstructured repeat [327]. The cleavage almost always results in an 8 nt sequence tag of repeat sequence that is at the 5' end of the mature crRNA [32, 99, 103, 129–131, 227, 274, 326]. The 3' ends of crRNAs are either cut to characteristic lengths by a largely unknown ruler mechanism [129], or part of the repeat remains at the 3' end [155]. The crRNA is stabilised by integration into its respective Cascade or Cmr complex to await the arrival of invader species. Type II systems are unique: processing is enabled by a *trans*-encoded RNA (tracrRNA) and the double-stranded RNA is processed by the endogenous RNaseIII [60]. In the final interference phase, the effector complex binds to the invading DNA (or RNA in some type III systems [295]) via base pairing between the crRNA and the protospacer. This interaction generally requires a seed interaction where near-perfect complementarity must exist proximal to the PAM motif, and looser complementarity can exist in more distal positions [145]. Illustration is adapted with permission from Macmillan Publishers Ltd: *Nature Reviews Microbiology* [202] copyright 2011 (license no. 3363730101764).

sufficient for targeting and cleaving foreign DNA. Due to its simplicity, the type II system has been adapted as a tool for genome editing and has been applied to a multitude of eukaryotes, including humans [308, 328]. Although the simplicity of the type II system has lead to its application in genetic engineering, the Cmr complex presents a specific advantage: it is the only known CRISPR complex to date that can target and degrade RNA instead of DNA. Thus, the Cmr complex of type III-B systems could be applied to post-transcriptional gene knock-down, similar to RNAi [308]. CRISPR loci can also be useful for differentiating between strains: active CRISPR loci uptake new spacers and undergo large mutations such that many strains of the same species can be differentiated by their CRISPR loci. A further application is to follow the history of genetic invasions into the prokaryote by mapping the spacers to sequenced invader species. This can be especially useful for tracking invading viruses into the prokaryote population in humans [256]. Current knowledge about this versatile system is still far from complete and the future looks bright for the rising CRISPR star.

2.3 Measuring prediction performances

One main task in bioinformatics is to develop tools that predict an outcome of a biological experiment. First, to compare the performances of different prediction approaches, we require training data (from which the prediction model can be learned) and test data (to which the prediction models are applied to compare performances). In this step, it is crucial that the training and testing data do not overlap and are independent. Now we assume that we have binary data divided into *positive* and *negative* instances. For example, let the instance describe an mRNA and a *trans* factor, then they either form a regulatory interaction (positive instance) or they do not form a regulatory interaction (negative instance). Predictions on a test dataset, where the nature of the instances is known beyond reasonable doubt, can be compared using equations based on the intersections between predictions and true observations defined in Table 2.1[1]. Various measures of prediction performances exist that consider different aspects. Selected measures are given in Definitions 2.1–2.5.

Table 2.1. Confusion matrix for dividing test data according to prediction outcomes.

		"Truth"	
		Positives ($\mathcal{P}^t$)	Negatives ($\mathcal{N}^t$)
Prediction	Positives ($\mathcal{P}$)	True Positives (TP)	False Positives (FP)
	Negatives ($\mathcal{N}$)	False Negatives (FN)	True Negatives (TN)

Definition 2.1. *The sensitivity (also known as recall) gives the proportion of positive instances that were correctly predicted as positive:*

$$\textit{sensitivity} = \frac{TP}{TP + FN} = \frac{TP}{\mathcal{P}^t}.$$

[1] False positives represent type I errors and false negatives represent type II errors.

Definition 2.2. *The* ***specificity*** *gives the proportion of negative instances that were correctly predicted as negative:*

$$specificity = \frac{TN}{FP+TN} = \frac{TN}{\mathcal{N}^t}.$$

Definition 2.3. ***Precision*** *is the proportion of positive predictions that are true positives:*

$$precision = \frac{TP}{TP+FP} = \frac{TP}{\mathcal{P}}.$$

Definition 2.4. *The* ***false discovery rate (FDR)*** *describes the proportion of positive predictions that are false:*

$$FDR = \frac{FP}{TP+FP} = \frac{FP}{\mathcal{P}}.$$

Definition 2.5. *The* ***accuracy*** *describes the proportion of all data instances that were correctly predicted as positive or negative:*

$$accuracy = \frac{TP+TN}{\mathcal{P}+\mathcal{N}}.$$

Many standard prediction tools not only output discrete (binary) predictions that would lead to single confusion matrix (Table 2.1), but produce probabilities, scores, or rankings. In these cases it is difficult to select a single threshold for a binary classification and changing the threshold would lead to different numbers of TP, FP, FN, and TN predictions. Thus, instead of just selecting an arbitrary setting, prediction performances can be measured more robustly by iterating over all settings and reporting the results. These can be visualised by receiver operating characteristic (ROC) curves [82] that plot the sensitivity (x-axis) as a function of the false discovery rate (y-axis)[1]. Reporting single measurements is done by computing the area under the ROC (called AUROC). In the ROC curve, assigning random predictions to the data instances would result in the diagonal line $y = x$ and this equates to an AUROC of 0.5. Any curve above the diagonal (corresponding to AUROC>0.5) points to a prediction performance better than random assignment). In general, results with AUROC$\geq$0.7 are recognised as convincing performances. However, ROC curves can be misleading when the numbers of positive and negative data instances are significantly different. In addition, sometimes the true extent of negative instances are unknown, especially in biological data. Instead of the ROC, it is possible to plot the recall (x-axis) as a function of the precision (y-axis) [82] and to report the area under this curve (AUPR). The AUPR concentrates only on the prediction of positive instances and disregards the prediction of negative instances.

[1] The performance of binary prediction tools can also be plotted in the ROC space, but as a single point rather than a curve.)

2.4 Definition and verification of RNA structure

The main purpose of DNA is to store the genetic information of an organism. The information it contains is communicated to the rest of the cell via the medium of RNA. Once transcribed, the RNA has multiple functions: (1) transferring protein-coding information to the ribosome, (2) regulating gene expression, and (3) catalysing biochemical reactions. This thesis deals with aspects of the (2) function. RNA-based regulation and interaction with binding partners is guided not only by affinity to its sequence, but its structure also plays a pivotal role (see Section 2.1.2 for examples). In the following, concepts, definitions and representations of RNA structure are established and approaches for experimental structure elucidation are briefly described.

2.4.1 Structure properties

RNA is a macromolecule comprising a chain of nucleotides that consist of three parts: a ribose sugar, a phosphate group, and a base (Figure 2.5). RNA structure is defined by bonds between bases and is influenced by further external conditions, such as temperature, salt concentrations, and availability of metal ions, the most important being Mg^{2+} ions. External conditions mainly influence the overall stability of an RNA structure, whereas the bases determine possible structure configurations. Since external influences are difficult to model computationally, prediction focusses on computing possible structure configurations.

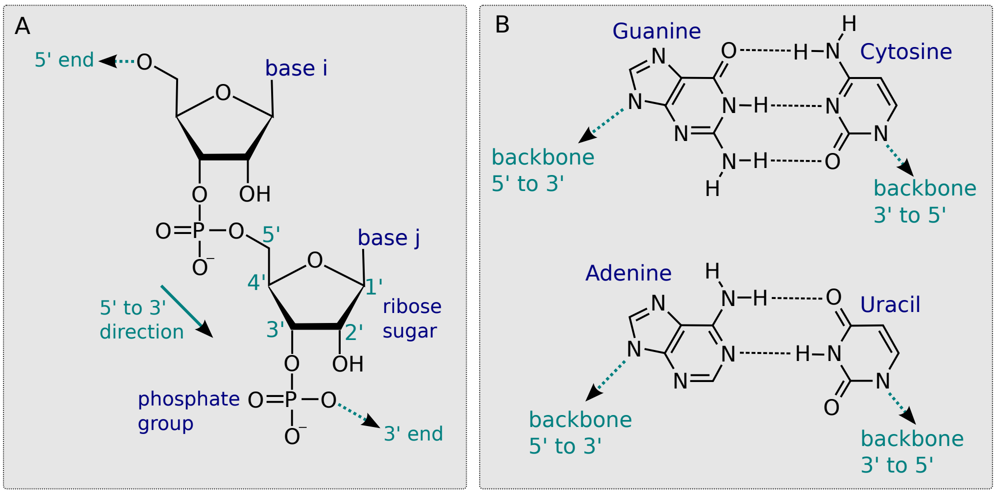


Figure 2.5. RNA molecules. (A) The RNA backbone is depicted with the alternating phosphate and ribose sugar groups and the orientation $5' \rightarrow 3'$, derived from the carbon-atom labelling, is visualised. (B) The four RNA bases—adenine *A*, cystosine *C*, guanine *G* and uracil *U*—are depicted with the Watson–Crick hydrogen bonds forming the most common base pairs: purines are on the left and pyrimidines on the right. The third-most-common base pair, *GU*, is not depicted, but forms two hydrogen bonds, each between an oxygen and an *N-H* group.

Primary structure

The primary structure (Definition 2.6) of an RNA molecule is defined by the order (i.e. sequence) of the bases in the molecule. RNA sequences are represented by their respective one-character symbols.

Definition 2.6. *The **primary structure** of an RNA consisting of* n *nucleotides is defined by the sequence* $R = (r_1, \ldots, r_n)$ *with* $r_i \in \{A, C, G, U\}$ *where* $A =$ *adenosine,* $C =$ *cytosine,* $G =$ *guanine and* $U =$ *uracil. All nucleotides* r_i *and* r_{i+1}*,* $\forall i \in \{1, \ldots, n\}$*, form the backbone of the RNA sequence, i.e., a covalent bond between the 3' end of* r_i *and the 5' end of* r_{i+1} *exists.*

RNA sequences are generally written in the 5' to 3' orientation (Figure 2.5.A). Visualisations of RNA should always indicate the 5' and 3' ends when it is unclear, since the orientation is important for biological processes, e.g., RNA synthesis always occurs in the 5' to 3' direction. Primary RNA sequences are frequently stored in the well-known FASTA format.

Secondary structure

Within an RNA molecule, two bases form hydrogen bonds between each other; a bonded pair of bases is called a base pair (Figure 2.5.B; Definition 2.7).

Definition 2.7. *A **base pair** in* R *of length* n *is a tuple* (i, j)*, such that* $(r_i, r_j) \in \{(G, C)$*,* (C, G)*,* (A, U)*,* (U, A)*,* (G, U)*,* $(U, G)\}$ *with* $r_i, r_j \in R$*.*

On occasion, it is necessary to know the distance of an intramolecular base pair, which is defined by the *bp-span* (Definition 2.8).

Definition 2.8. *The **base-pair span** defines the distance between two nucleotides* $1 \leq i < j \leq n$ *with respect to their position on the RNA sequence: bp-span*$(i, j) = j - i + 1$*.*

The most common definition of the secondary structure is defined by the set of non-crossing (nested or adjacent) base pairs in an RNA sequence. The easiest way to define the RNA secondary structure is as a graph (Definition 2.9); the last two conditions (5) and (6) below do not have to be met, but are common assumptions in prediction algorithms and they hold for all single-RNA secondary structures in this thesis.

Definition 2.9. *The **secondary structure** of* R *of length* n *is an undirected graph* $S = (N, B)$ *with* $N = \{1, ..., n\}$ *the set of nucleotides and* $B \subset N \times N$ *the set of bonds between nucleotides, such that*

1. $(i, j) \in B$ *and* $(j, i) \in B$ *(graph is undirected),*

2. $(i, i + 1) \in B, \forall i, 1 \leq i < n$ *(represents RNA backbone),*

3. $(i, i) \notin B, \forall i \in N$ *(no self loops),*

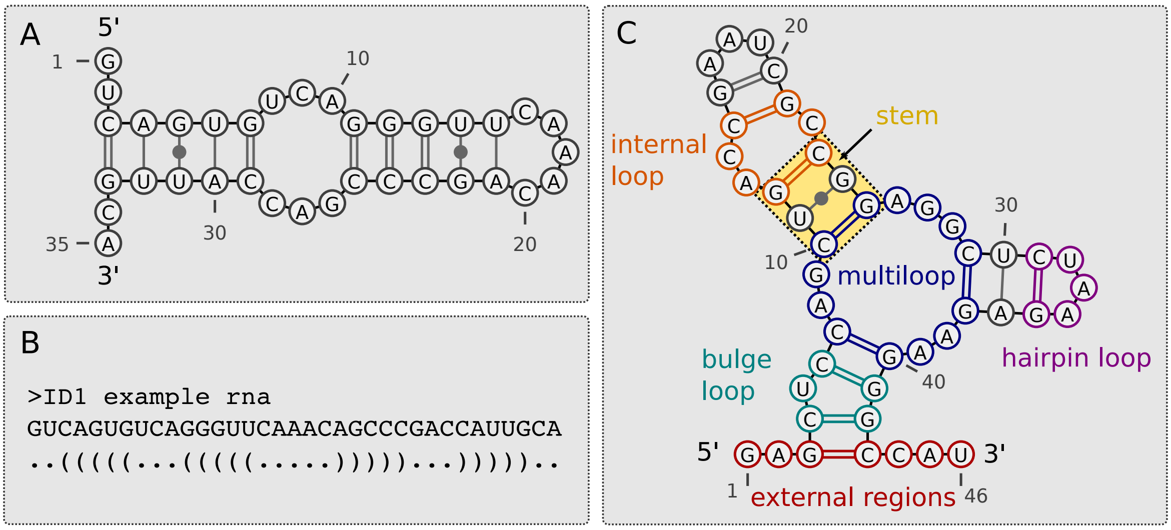
4. *$\forall i \in N$, there exists at most one $j \neq i \pm 1$ with $(i, j) \in B$ (a base can be paired to at most one other),*

5. *let (i, j) and (k, l) be two base pairs, then $i < k < l < j$ (nested base pair) or $i < j < k < l$ (adjacent base pair) is true $\forall (i, j), (k, l) \in B$ (no crossing base pairs),*

6. *$\forall (i, j) \in B$, bp-span ≥ 5 (at least 3 unpaired bases are required for the RNA to turn back on itself),*

7. *if $(i, j) \in B$ then $(i - 1, j + 1) \in B$ or/and $(i + 1, j - 1) \in B$ (no lonely base pairs).*

A base pair in the RNA secondary structure $S = (N, B)$ is represented by an undirected edge $(i, j) \in B$, such that $(j, i) \in B$ (condition (1) in Definition 2.9). To reduce double entries for a single base pair, edges are always notated as (i, j) with $i < j$. In text, a specific base pair type with an unspecified direction is written as GC. Although, formally, a base pair is an undirected edge, the direction with respect to the $5' \rightarrow 3'$ orientation of the RNA sequence (Figure 2.6.A) can be biologically relevant. In this case, the direction is indicated by $C \rightarrow G$, whereby C is closer to the 5' end and G is closer to the 3' end of the RNA sequence. Several representation possibilities for RNA structures exist, however, for the sake of brevity, only those representations in Figure 2.6.A–B are used in this dissertation.



```
>ID1 example rna
GUCAGUGUCAGGGUUCAAACAGCCCGACCAUUGCA
..(((((...(((((.....)))))...)))))..
```

Figure 2.6. Secondary structure representations and elements. (A) The secondary structure represented in a planar graph layout for easy-to-understand visualisation. (B) The secondary structure in dot-bracket format, saved in an extended FASTA file format. Matching parentheses correspond to base pairs and the dots to unpaired bases. A FASTA file without structure information would not include the dot-bracket string. The dot-bracket structure format is both human and machine readable, thus the preferred format for bioinformatics tools. (C) Secondary structure elements visualised on an example in the graph layout: hairpin loop, bulge loop, internal loop, multiloop, external region multiloop, and stem. The bordering base pairs define the outer limits of the structure element, such that these elements overlap, that is two elements can share a base pair.

Secondary structure elements

A secondary structure can be broken down into several structure elements (Figure 2.6.C). For computational reasons, it is easy to define the elements by their bordering base pairs, however, in the case of loops and external regions, sometimes only the unpaired bases are important for biological processes, for example, if a *trans* factor only binds to single-stranded regions, it would only bind to the unpaired bases of a hairpin loop and not to the enclosing base pair. The number of consecutive unpaired bases in a loop determines the loop size.

When describing the general structure characteristics at binding sites of a particular *trans* factor, the terms **single-stranded** and **double-stranded** regions are frequently used. A single-stranded (or unpaired) region is a consecutive stretch of bases that do not pair with any other base (Definition 2.10). A double-stranded (or paired) region is the inverse, i.e., a stretch of consecutive nucleotides for which all bases form a base pair with any other base.

Definition 2.10. *Let $R = (r_1, \dots, r_n)$ be an RNA sequence with its structure conformation $S = (N, B)$. The interval $[x, y]$ for $1 \leq x < y \leq n$ is* ***unpaired*** *or* ***single-stranded*** *if $(i, j) \notin B, \forall i \in [x, y]$ and $\forall j \in \{1, \dots, n\}$.*

More detailed information about which type of structures a *trans* factor binds to is given by separating structure context into the different structure elements given in the subsequent Definition 2.11. Structures that consist of only stems and loops are often referred to as stem-loop structures, i.e. they do not contain multiloops.

Definition 2.11. *The secondary structure elements of RNA sequence $R = (r_1, \dots, r_n)$ with structure $S = (N, B)$ are defined below.*

- ***Hairpin loop:*** *is enclosed by a base pair $\mathcal{H} = (i, j) \in B, i < j$ where the interval $]i, j[$ is unpaired.*
- ***Internal loop:*** *is enclosed by two base pairs $\mathcal{I} = \{(i, j), (k, l)\} \subset B$ with $i < k < l < j$ and the intervals $]i, k[$ and $]l, j[$ are unpaired with $k - i > 1$ and $j - l > 1$.*
- ***Bulge loop:*** *is enclosed by two base pairs $\mathcal{B} = \{(i, j), (k, l)\} \subset B$ with $i < k < l < j$ and either $k - i > 1, j - l = 1$ and $]i, k[$ is unpaired or $k - i = 1, j - l > 1$ and $]l, j[$ is unpaired.*
- ***Multiloop:*** *is enclosed by at least 3 base pairs $\mathcal{M} = \{(i_1, j_1), (i_2, j_2), \dots, (i_m, j_m)\} \subset B$ and $\forall (i', j') \in \mathcal{M}, i' < j'$. The first base pair (i_1, j_1) is termed the closing base pair with $i_1 < i_2$ and $j_1 > j_m$. For all other base pairs $(i_q, j_q) \in M, \forall q, 2 < q < m, j_{q-1} < i_q < j_q < i_{q+1}$. All intervals $]i_1, i_2[,]j_m, j_1[$ and $]j_q, i_{q+1}[, \forall q, 1 < q < m$, are either empty or unpaired. A multiloop is short for a multi-branched loop and the number of 'branches' is given by $m - 1$.*
- ***External region:*** *is an unpaired interval $[e, f]$ in R where there exists no $(i, j) \in B$ for which $1 \leq i < e < f < j \leq n$ is true and $[e, f]$ is maximal in the sense that $e = 1$ or $(e', e - 1) \in B, 1 \leq e' < e$ and $f = n$ or $(f + 1, f') \in B, f < f' \leq n$.*

- ***All external regions:*** *are given by base pairs* $\mathcal{X} = \{(i_1, j_1), \ldots, (i_x, j_x)\} \subset B$*, such that all intervals* $[1, i_1[,\ ]j_x, n]$ *and* $]j_q, i_{q+1}[, \forall q, 1 \leq q < x$*, that contain at least one nucleotide are external regions. If no base pairs exist, then there is only one external region, given by* $[1, n]$*.*

- ***Stacking base pairs:*** *are two base pairs* $(i, j), (k, l) \in B$ *with* $i < k < l < j$ *such that* $k - i = 1$ *and* $j - l = 1$*.*

- ***Stem:*** *is defined by at least two base pairs* $\mathcal{T} = \{(i_1, j_1), \ldots, (i_t, j_t)\} \subset B$ *where* $\forall (i_q, j_q), (i_{q+1}, j_{q+1}) \in \mathcal{T}, (i_q, j_q)$ *and* (i_{q+1}, j_{q+1}) *are stacking base pairs.*

Tertiary structure

The tertiary structure is the one that drives the biological function and—as in proteins—is defined by the exact position of each atom in three-dimensional space. Here, the secondary structure elements are further stabilised by several van der Waals connections, additional hydrogen bonds and entropic factors.

There are three extensions to the base-pairing rules of secondary structures that are allowed in tertiary structures. The first includes base pairs that are different from the regular ones in Definition 2.7 [188]. The second extension allows pseudoknots. A pseudoknots describes a secondary structure where crossing base pairs (Definition 2.12) are allowed. Third, the rule of one base pairing with at most one other can be broken [52]. On occasion these extensions to the base-pair set are considered to also be specialised secondary structures. This happens when specialised prediction approaches consider such extended base pairing, but still ignore the exact positioning of the atoms in three dimensions [27,48,73,182,219,244,253,254,284]. All possible extensions to the secondary structure and the final tertiary structures are not considered in this thesis, therefore, only a brief overview suffices at this point.

Definition 2.12. *Let* $(i, j) \in B, i < j$*, then* $(k, l) \in B, k < l$*, is a* ***crossing base pair*** *if* $i < k < j < l$ *or* $k < i < l < j$*.*

2.4.2 Experimental verification

A popular approach to elucidating RNA structure involves enzymatic or chemical probing [232]. The probe measures the reactivity of single nucleotides to a specific enzyme or chemical. Depending on the properties of the enzyme or chemical, structure propensities of that nucleotide can be deduced. Probes specifically cleave or chemically modify nucleotides that are either bound or unbound and usually complementary probes are used so that a ratio of paired vs. unpaired bases can be determined. For example, RNase T1 cleaves specifically at unpaired guanines and RNase V1 cleaves double-stranded regions[1]. The chemical reagents used the SHAPE probing technique [198,223,301] react with the RNA backbone, probing its

[1] RNase V1 is also known to cleave stacked, but not paired, nucleotides [232].

mobility, such that all four nucleotide types can be probed in a single experiment. Combined with high-throughput sequencing, chemical or enzymatic probing can be used to determine structure characteristics on a transcriptome-wide scale [71,165,198,268,278,301,317,336].

The downside of the structure probing approach is that exact base pairing cannot be determined. Thus, an extension to structure probing is to test the effect of mutations that destroy or extend putative secondary structures. Subsequently, the structure probing is repeated, or in functional studies, a previous observation is either prevalent or absent after mutation, e.g. in [129,227,298]. Mutational studies cannot, however, be performed in high throughput. Therefore, computational approaches exist that incorporate structure probing results into secondary-structure–prediction algorithms [59,237,334].

Tertiary structures can be determined via X-ray cristallography or NMR spectroscopy [261,280]. Structures are deposited in databases, such the Protein Data Bank, which also incorporates RNA structures [127], and `RNA STRAND` [7]. Tertiary-structure determination is both time and cost expensive and currently not suited to large-scale analyses.

2.5 RNA-structure prediction approaches

A major advantage of RNA structure prediction in contrast to protein-structure prediction is that, in general, the secondary structure contributes substantially to the free energy of the final tertiary structure [289,312]. Thus, RNA structure can be approximated by concentrating on the prediction of secondary structure: an observation that has been exploited by most RNA-structure prediction algorithms.

2.5.1 The nearest-neighbour energy model

Structure prediction, based on thermodynamics, requires an energy function $\mathrm{E} : S \to \mathbb{R}$ to evaluate the potential for an RNA sequence R to fold into a given structure S in aqueous solution and ultimately in the cell. To this end, current secondary-structure prediction approaches base their algorithms on the assumption that the *change in the Gibbs free energy* ΔG of a fixed structural conformation reflects its folding potential. The change in Gibbs free energy is equal to the work exchanged between the RNA molecule with its surroundings, depending on pressure and temperature forces, during the reversible process of RNA folding. A $\Delta G < 0$ indicates a favourable process in which the folded RNA is stabilised relative to the unfolded form. The lower the ΔG value, the more stable the structure is; thus, the lowest possible ΔG for R is assumed to be optimal. However, this assumption does not take time into account. The biologically functional structure might not possess the optimal (i.e. minimum) change in free energy, but could be a structure that forms more quickly (in terms of time) and is thus more relevant for functioning in its cellular surroundings [88]. Therefore, suboptimal stable structures must also be considered.

Current secondary structure prediction approaches use the *nearest-neighbour energy model* to estimate free energies. This model requires a decomposition of an RNA structure into a set of basic structure units that overlap with a neighbouring element by one base-pair. These structure units are analogous to the loop elements defined in Definition 2.11 in Section 2.4.1 where the stems are further decomposed into two stacking (consecutive) base pairs. The energy contributions of these basic elements can be measured experimentally [315]. These measurements provide energy parameters that are applied in a structure prediction model. Multiple sets of such energy parameters exist [207, 208, 315] and have been improved over time; the selection of the energy parameters is a very important aspect that will change prediction results.

The nearest-neighbour energy model assumes that the overall change in energy of an RNA structure is equal to the sum of the energy contributions of all the basic structures in its decomposition ΔG (Definition 2.13). It is termed such, because the energy contribution of a base pair, for example, is only dependent on the next base pair, i.e., on its *nearest neighbour*. It is possible to have dependencies that are structurally more distant, but these are ignored for the sake of simplicity and computability. Further dependencies would require more experimental measurements, which has not yet been feasible.

Definition 2.13. *Let $S = \{s_1, \ldots, s_m\}$ be the decomposition of RNA structure S into its basic substructures s_i. The **energy function** of the **nearest-neighbour-energy model** is given by*

$$E(S) = \sum_{i=1}^{m} e(s_i) \approx \Delta G(S),$$

where $e(s_i)$ is the measured energy contribution for the substructure s_i and $\Delta G(S)$ is the change in Gibbs free energy for S.

2.5.2 The optimal structure and base-pair probabilities

The structure with the lowest free energy, i.e., the minimum-free-energy (MFE) structure (Definition 2.14), is assumed to be optimal. Although this structure is not always biologically functional, the probability that it will form (at equilibrium over time) is high—if the energy model is correct and the activation energy is not too high.

Definition 2.14. *Let $\mathcal{Q}_R = \{S_i, \ldots, S_r\}$ represent the ensemble of all possible secondary structure configurations of RNA sequence R. The **minimum-free-energy (MFE) structure** of R is $S_i \in \mathcal{Q}_R$ where $E(S_i) \leq E(S_k), \forall S_k \in \mathcal{Q}_R$.*

To calculate the MFE structure, one has to evaluate the energy of all possible structure configurations of the RNA sequence and identify the one with the lowest energy. Since identifying the MFE structure according to the nearest-neighbour energy model satisfies the Bellman's principle of optimality[1], the energies of possible structures can be evaluated (according to the minimum) efficiently using a dynamic programming recursion and the

[1] "*An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision*"—Richard Bellman 1957.

MFE structure can be identified using a trace back in the dynamic programming table; this algorithm was first introduced by Michael Zuker and Patrick Stiegler in 1981 [356] and is referred to as the “Zuker algorithm”. The time complexity of the Zuker algorithm for an RNA of length n is $\mathcal{O}(n^4)$, which can be reduced to $\mathcal{O}(n^3)$ if the loop size[1] is set to a maximum (usually 30 nt).

As previously established, the output of the Zuker algorithm may not coincide with the biologically functional structure. There are many reasons for wanting to identify multiple highly probable structures: (1) for every RNA sequence, more than one MFE structure might exist, however, only one MFE structure is given as the output of the algorithms; (2) sequences can switch between structure variants and these variants can be functionally relevant, for example, riboswitches alternate between an *on* and an *off* structure in regulatory processes [29]; (3) the functional structure may form via a kinetic as opposed to a thermodynamic pathway at equilibrium [88]; and (4) the energy model and energy parameters only estimate the real ΔG. Furthermore, interacting molecules are not taken into account, therefore, the calculated MFE structure may not be the true MFE structure. Thus, it is often more informative to consider the entire ensemble of structures and to predict structure probabilities.

In 1990, J. S. McCaskill introduced an algorithm that employs the partition function on the Boltzmann-distributed ensemble of all possible structure configurations for a single sequence (Definition 2.15) to calculate the probability of a given RNA structure or base pair [210].

Definition 2.15. *Let $\mathcal{Q}_R$ be the ensemble of all possible secondary structure configurations of RNA sequence R. The* ***partition function*** *of $\mathcal{Q}_R$ is defined as:*

$$Z(\mathcal{Q}_R) = \sum_{S \in \mathcal{Q}_R} e^{-\frac{E(S)}{\mathcal{R}\mathcal{T}}},$$

where $\mathcal{R} = 8.3146$ is the gas constant in joules per degree Kelvin and $\mathcal{T}$ is the absolute temperature in degrees Kelvin.

The total energy of the structure ensemble is often used as a measurement for the *structuredness* of an RNA sequence (Definition 2.16): the lower the ensemble energy, the more stable the structures, in general, that are formed by the respective RNA—considering all possible structure configurations.

Definition 2.16. *The* ***ensemble energy*** *of an RNA sequence R is given by:*

$$E(\mathcal{Q}_R) = -\mathcal{R}\mathcal{T} \ln Z(\mathcal{Q}_R),$$

where $Z(\mathcal{Q}_R)$ is the partition function over the ensemble of all structures $\mathcal{Q}_R$. The ensemble energy is used to measure the overall ***structuredness*** *of R.*

The probability of observing a certain structure S in the structure ensemble $\mathcal{Q}_R$ is given by the Boltzmann-weighted structure energy, divided by the partition function of the structure ensemble:

[1] The loop size is determined by the number of consecutive unpaired nucleotides in a loop.

$$Pr[S|R] = \frac{e^{-\frac{E(S)}{\mathcal{R}\mathcal{T}}}}{Z(\mathcal{Q}_R)}. \tag{2.1}$$

Following this calculation, a *base-pair probability* is simply defined as the partition function over all structures that contain the base pair (i, j), divided by the partition function of the entire structure ensemble:

$$p(i, j) = \frac{Z(Q_{(i,j)})}{Z(\mathcal{Q}_R)}, \tag{2.2}$$

where $Q_{(i,j)} \subset \mathcal{Q}$ is the set of structures that contain the base pair (i, j). Base-pair probabilities can be calculated similarly to the MFE structure in $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space[1].

The Zuker and/or the McCaskill are implemented by `RNAfold` [138], `UNAfold` [206], `Rfold` [170] and `RNAstructure` [255]. Although they are based on the same underlying algorithms, they differ by some additional options that can aid with structure analysis. All structure predictions performed in this work are done using the RNA Vienna Package that includes `RNAfold`.

2.5.3 Finding functional structures in the suboptimal space

MFE-structure and base-pair prediction algorithms are the traditional approaches to the structure prediction of single RNA sequences. Although base-pair probabilities give a good estimation of the whole ensemble of structures and which base pairs are more probable than others, it does not output highly likely, exact secondary-structure configurations. Alternatively to reporting the MFE structure, it is possible to calculate the centroid structure [68], which is the best representative of all structures in the Boltzmann-weighted ensemble. The centroid structure has been reported to be more accurate than the MFE structure [68], however, it is still prone to errors and does not solve the problem of multiple biologically active structures. One solution to gaining multiple, probable structure configurations is merely listing all structures with sub-optimal Gibbs free energies, as is done in `RNAsubopt` [344]. An energy cutoff or number of top structures is given as a parameter. However, since every base-pair change is considered a different structure, many interesting structures are very far down that list and it is difficult to find biologically relevant structures. An alternative is provided by `RNAshapes` [296]: here single base-pair changes are ignored, but structures are categorised into groups according to their general shape. For example, you can group your output into structures that resemble a single stem-loop structure, irrelevant of the size of the hairpin, internal, and bulge loops, or their frequency—or all structures that resemble the classic tRNA three-branched multiloop could be grouped together. The number and size of shape categories are regularised by five shape extraction levels. The first level only differentiates between stem-loops and the number of branches in multiloops; the fifth level is

[1] Details of the algorithm, i.e., the dynamic programming recursions, are not required for understanding the work presented in this thesis and are thus omitted; they can be taken from the original publications or related literature.

extended to differentiating between the number of internal and bulge loops. Most importantly, the shape extraction always disregards the length of a stacking region. The structure in each shape category with the minimimum free energy is called the *shrep* structure and is reported. These shreps allow for a better overview of the structure space and perhaps an easier identification of biologically active structures. Further approaches use statistical sampling from the Boltzmann-probability–weighted structure ensemble to report a moderately-sized list of viable structures [67, 70].

2.5.4 Local structure prediction

The Zucker and McCaskill algorithms are considered to perform a *global* structure prediction: structures are computed for the entire input RNA sequence and all possible base pairs are allowed. These methods are well suited to predicting structures of short regulatory non-coding RNA that form a global structure, for example miRNA precursors or tRNAs (Section 2.1.2). In contrast to the global approach, a *local* structure prediction would only compute structures that are local in the sense that base pairs only span a subsequence of the RNA. The motivation for developing local structure prediction methods is both biological and computational. First, RNA-based regulation is not only guided by ncRNAs that form global structures, but often longer RNA species contain local structure motifs that are important for *trans*-factor binding. Second, the cubic time and space complexity of global structure prediction makes their application to very long RNAs (mRNAs can span many kilobases) unfeasible.

A first algorithmic solution to the high runtime complexity was to limit the distance on the sequence between two base pairs, i.e., the base-pair span (see Definition 2.8) and to ignore any base pairs with spans larger than a given threshold, typically denoted by L. Let $\mathcal{Q}^L$ be the set of structures possible for R that have a maximum base-pair span of L. Then the probability of these more local base pairs is:

$$p^L(i,j) = \frac{Z(\mathcal{Q}^L_{(i,j)})}{Z(\mathcal{Q}^L)}, \tag{2.3}$$

where $\mathcal{Q}^L_{(i,j)} \in \mathcal{Q}^L$ is again the set of structures that contain the base pair (i, j) and $Z(\mathcal{Q})$ is the partition function over the set of all structures in $\mathcal{Q}$. As this approach still folds the entire input sequence simultaneously and merely restricts the base-pair spans of the predicted structures, it can be considered as *semi-local*. Implementations are `RNALfold` [139] to find locally stable structures and `Rfold` [170] for base-pair probabilities.

The second algorithmic solution was to predict structures in sliding windows of a fixed length denoted by W (Definition 2.17), in addition to the maximum base-pair span constraint L and $W \geq L$.

Definition 2.17. *A* ***window*** $\mathcal{W}^u$ ***of length*** W *is defined by an interval* $\mathcal{W}^u = [u, u+W-1]$*, where* $(r_u, \ldots, r_{u+W-1})$ *is a subsequence of* R*.*

The probability of a base pair within the window $\mathcal{W}^u$ is:

$$p^{\mathcal{W}^u,L}(i,j) = \frac{Z(\mathcal{Q}^{\mathcal{W}^u,L}_{(i,j)})}{Z(\mathcal{Q}^{\mathcal{W}^u,L})},$$

where $\mathcal{Q}^{\mathcal{W}^u,L}$ is the set of all possible structures for the window $\mathcal{W}^u$ and $\mathit{bp\text{-}span}(i',j') \leq L$ for all (i',j') in structures of $\mathcal{Q}^{\mathcal{W}^u,L}$, then $\mathcal{Q}^{\mathcal{W}^u,L}_{(i,j)} \subset \mathcal{Q}^{\mathcal{W}^u,L}$ is the subset of structures that contain the base pair (i,j). Now, in the sliding-window approach, a base-pair probability is averaged across all windows that it occurs in:

$$p^{L,W}_{avg}(i,j) = \frac{1}{W - \mathit{bp\text{-}span}(i,j)} \cdot \sum_{u=j-W+1}^{i} p^{\mathcal{W}^u,L}(i,j). \tag{2.4}$$

The average base-pair calculation allows for a single score for all base pairs in the entire input sequence R. Note that $p^{L,W}_{avg}(i,j)$ is not a probability, but represents the normalised expected number of base-pair occurrences over all windows. This *window-based* approach is *local* in the sense that each window is folded independently of the rest of the sequence[1]. Approaches that predict true local structures, without the use of fixed windows, currently do not exist. The window-based approach is implemented in `RNAplfold` [18, 19].

An RNA sequence of length n can be now be computed in $\mathcal{O}(nL^2)$ time and $\mathcal{O}(n + L^2)$ space, which is basically linear for L values that deliver accurate results (see Part IV).

2.5.5 Dotplots: Visualising base-pair probabilities

Prediction results in the form of base-pair probabilities are presented as **dotplots**. Base-pair probabilities are visualised as a square matrix D with cells $D_{i,j}$ such that $r_i, r_j \in R$. The RNA sequence R is written along the sides of the matrix (in 5'→3' orientation from top to bottom and left to right). Dots in upper-right cells ($D_{i,j}, i < j$) represent base pair probabilities, whereas, dots lower-left cells ($D_{i,j}, i > j$) represent base pairs involved in the MFE structure. The size of a dot (in upper-right cells) is proportional to the base-pair probability. Local structure prediction results can also be visualised as dotplots, but in this case, only the top right triangle is depicted (rotated by −45°) and cells where the $\mathit{bp\text{-}span} \geq L$ are omitted.

2.5.6 Accessibility

Accessibility is a term used to describe how "unpaired" a segment of RNA is to determine whether it is "accessible" for binding by regulatory factors. It is commonly measured as either the free energy required to reverse the formation of any base-pairs within the RNA segment of interest so that it becomes single stranded; or as the probability of that segment to be unpaired (i.e. single-stranded). We use probabilities to measure accessibility in this work, which usually makes sense for short segments. Probabilities become too small for

[1] Although it is possible to set $L = W$, we show in Chapter 6 that when $L \ll W$ detrimental effects of the artificial window borders, introduced by the sliding windows, can be avoided.

longer RNA segments in which case energies are preferred. In many instances, it has been shown that the target site in the RNA—to which a *trans* factor binds—should be accessible for binding: some data are available that indicate that around miRNA target sites the accessibility is significantly increased in comparison with random contexts or non-target sites [143, 164, 171]. The same is true for siRNAs [109, 303]; and many RNA-binding proteins bind to single-stranded regions [249], e.g., the splicing factor SRSF1 [329]. This means that base-pairing within the target region would reduce the *accessibility* of the binding site. Computation of accessibility is thus important and can be computed in a similar way to base-pair probabilities. The position-wise accessibility $pu(i)$ is the probability of base r_i in the RNA sequence R being unpaired. Hence, the accessibility of r_i is the probability of complementary event of r_i being paired, which can be derived from the sum of all base-pair probabilities involving r_i:

$$pu(i) = 1 - \sum_{j=1}^{n} p(i,j), \tag{2.5}$$

where $p(i,j)$ is the probability for the base-pair (i,j).

In regulatory mechanisms, binding of a *trans* factor usually requires a stretch of nucleotides to be unpaired and not just a single nucleotide. The simplest solution is to calculate the average unpaired probability of single nucleotides for the region of interest. However, it is also possible to calculate the probability that the interval $[v,w]$ is unpaired (i.e., single-stranded), analogously to the probability of a base-pair (see Equation 2.2):

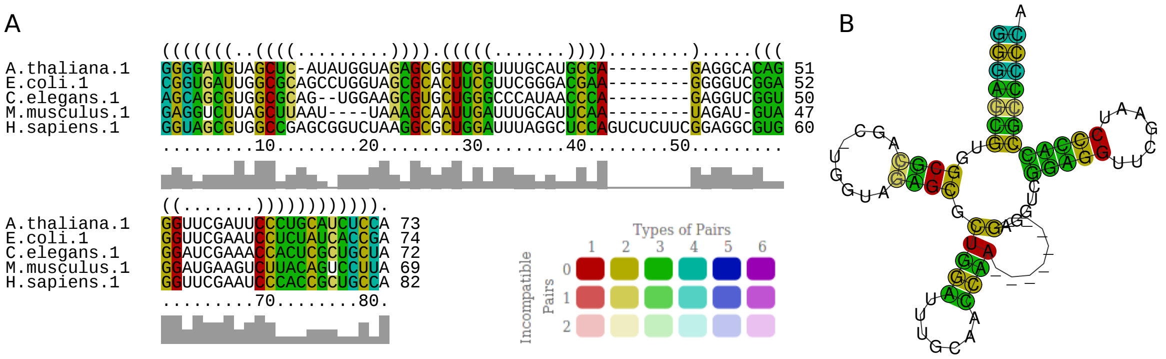
$$pu(v,w) = \frac{Z(\mathcal{Q}_{[v,w]})}{Z(\mathcal{Q}_R)}, \tag{2.6}$$

where $Z(\mathcal{Q}_{[v,w]})$ is the partition function (Definition 2.15) over all structures for which the interval $[v,w]$ is unpaired (Definition 2.10) and $Z(\mathcal{Q}_R)$ is the partition function over all possible structures for sequence R. Since many applications of accessibility are on long sequences, such as mRNAs, the program `RNAplfold` offers the calculation of the accessibility of all possible intervals up to a maximum size, averaging $pu(v,w)$ over all windows that include the interval $[v,w]$ (analogously to the average base-pair probabilities in Section 2.5.4) [18, 19]. We denote this mean probability as $\overline{pu}(v,w)$ and it represents the normalised expected frequency that the interval $[v,w]$ is accessible over all windows.

As already mentioned, accessibility is also measured as the cost (the energy required) of opening base pairs at the binding site—the interval $[v,w]$—written as δG_{open} [164] or ED [34]. The opening energy can be computed directly from the unpaired probability: $\delta G_{open}(v,w) = -\mathcal{R}\mathcal{T} \ln pu(v,w)$, where $\mathcal{R}$ is the gas constant and $\mathcal{T}$ the absolute temperature, and $pu(v,w)$ is often approximated by $pu_{avg}(v,w)$, which is the averaged probability as computed by the window-based prediction approach, `RNAplfold` [18, 19].

2.6 Use of conservation to detect non-coding RNA

As previously established, structure is generally paramount to RNA-based regulation. In many regulatory RNAs, the global structure of homologs is conserved, whereas the sequence is only conserved in small, local subregions of the RNAs (see Figure 2.7). When considering a secondary RNA structure, mutating a $G \to C$ base pair to first $G \to U$ and then $A \to U$ means that although the sequence differs, the base pair—the overall structure and ultimately the function—is conserved; such events are referred to as *compensatory base-pair mutations*. If several compensatory base-pair mutations occur over time, the sequence divergence can be great with comparable structure conformations (e.g. see Figure 2.7).



```
              (((((((..((((.........)))).(((((.......))))........).....(((
A.thaliana.1  GGGGAUGUAGCUC-AUAUGGUAGAGCGCUCGCUUUGCAUGCGA--------GAGGCACAG 51
E.coli.1      CGGUGAUUGGCGCAGCCUGGUAGCGCACUUCGUUCGGGACGAA--------GGGGUCGGA 52
C.elegans.1   AGCAGCGUGGCGCAG--UGGAAGCGUGCUGGGCCCAUAACCCA--------GAGGUCGGU 50
M.musculus.1  GAGGUCUUAGCUUAAU----UAAAGCAAUUGAUUUGCAUUCAA--------UAGAU-GUA 47
H.sapiens.1   GGUAGCGUGGCCGAGCGGUCUAAGGCGCUGGAUUUAGGCUCCAGUCUCUUCGGAGGCGUG 60
              .........10........20........30........40........50.........

              ((.......)))))))))))).
A.thaliana.1  GGUUCGAUUCCCUGCAUCUCCA 73
E.coli.1      GGUUCGAAUCCUCUAUCACCGA 74
C.elegans.1   GGAUCGAAACCACUCGCUGCUA 72
M.musculus.1  GGAUGAAGUCUUACAGUCCUUA 69
H.sapiens.1   GGUUCGAAUCCCACCGCUGCCA 82
              .........70........80.
```

Figure 2.7. Multiple sequence-and-structure alignment of tRNAs. Selected tRNAs from various model organisms, taken from `Rfam` [33, 95] (RF00005), were aligned using `LocARNA` [291, 339]. The resulting sequence–structure alignment with columns coloured according to the number of compensatory base-pair mutations as indicated in the legend is shown in (A); the sequence conservation per column is given by the grey bars. The characteristic tRNA cloverleaf structure (B) is highly conserved as seen again by the colouring as in the alignment. Only very few base-pairs are red, i.e., have identical sequences in all five organisms. With this example, we see that a common structure configuration is more important to the tRNA function (with up to 4 out of 5 possible compensatory base-pair types) than conserved sequence: the mean pairwise sequence identity is 53 %, which is still fairly high for an ncRNA class.

In all areas of bioinformatics, predicted functional annotations rely heavily on the use of conservation. If a signal is observed (or similar) in many species, especially if they are distantly related, then the common assumption is that the signal derives from an important function due to the constraints on random mutations that occur throughout evolution. Thus, the conservation of RNA structure is a powerful tool for detecting ncRNA—if the sequence similarity is too low, structure conservation provides an additional layer of information. The first step in identifying classes of ncRNA is to accurately predict alignments for multiple sequences that represent potential homologs of ncRNA across many species. Algorithms for predicting such alignments should consider both sequence conservation and detect compensatory base-pair mutations that conserve the global structure. A brief summary of these approaches is given in the following section.

The annotation and detection of novel ncRNAs in entire genomes requires whole genome alignments from a set of species with sufficient—but not too much—sequence homology, and second, a process for recognising domains within the genome that show significant structure conservation. Since ncRNA genes do not code for proteins, one can not look for characteristic open reading frames and signatures of codon usage. Instead, conserved structures is often the only chance for *in-silico* detection. Several sliding-window–based approaches, such as RNAz [118,333], EvoFold [240], and REAPR [341], have been developed specifically for ncRNA detection.

2.6.1 RNA sequence-and-structure alignment

Given a set of homologous RNA sequences, the task is to find an alignment that best reflects evolution of the sequences: conserved nucleotides should be aligned in single columns and conserved base pairs aligned in respective pairs of columns. Accurate alignments give rise to a conserved structure that is essential for the function of an ncRNA family. Two popular approaches for computing sequence-and-structure alignments exist. The first approach requires a multiple-sequence alignment [76,160,231,311] to be precomputed and a consensus structure is predicted by considering the columns of the precomputed alignment. The MFE consensus structure is computed by combining the information from compensatory base-pair mutations with a standard dynamic-programming, structure-folding algorithm (as described in Section 2.5); various implementations of this nature are RNAalifold [17], pfold [173] and PETFOLD [281]. This sequence-then-structure approach works well when the average pairwise sequence identity is ≥ 60 % [96,332]. The second main approach computes alignments where sequence and structure conservation are considered simultaneously; it is more accurate on the many ncRNA families with low pairwise sequence identities [96,133]. The first algorithm was introduced by Sankoff in 1985 [273]. It compares all possible structure configurations of one sequence with all possible structure configurations of the second sequence in a Needleman–Wunsch-like alignment of two sequences [229]. The Sankoff algorithm is not practical for more than two or long sequences because k sequences of length n take $\mathcal{O}(n^{3k})$ time and $\mathcal{O}(n^{2k})$ space. Many heuristics have appeared in recent years that restrict the Sankoff algorithm to a light-weight version and only perform pairwise structure alignments. Multiple sequences are subsequently aligned using the Feng–Doolittle progressive, or an iterative approach, analogous to sequence-only alignments [84]. Among the many heuristics, PMcomp [137] and LocARNA [339] introduced the use of precomputed base-pair probability matrices for increased computing efficiency. They are, however, still too slow for whole genome analyses with a time complexity of $\mathcal{O}(k^2n^4)$. Throughout this thesis, no long, or large-scale sets of RNA were aligned; therefore, due to its overall accurate performance and easy-to-use web server [291], LocARNA was used throughout this thesis.

2.6.2 Families of non-coding RNA

RNA families are stored in the Rfam database [33, 95]: version 11.0, released in August 2012, comprises 2208 conserved families of structured regulatory RNA—including ncRNA genes, self-splicing RNAs, and local structured *cis*-regulatory elements. Once an ncRNA family has been established, and accurate multiple sequence-and-structure alignments of that family exist, this information can be deployed for identifying further family members. The most basic and easiest approach for identifying family members is to use `BLAST` [5]—and in many cases this works if sequence similarity is sufficient. A more advanced approach is to include RNA secondary structure. For example, `INFERNAL` [228] computes probabilistic covariance models of both RNA structure and sequence variation, capturing compensatory base-pair mutations, from precomputed multiple sequence alignments. This information is encoded by stochastic context-free grammars, which are extensions of hidden Markov models used for protein families [86], that can cope with the long-range base-pair interactions. Covariance models of established ncRNA families can be applied to scan for additional family members [349]. `LocARNA-SCAN` [340] offers an alternative to `INFERNAL` and gives a good comparison of sequence-only and sequence-and-structure–based approaches.

2.7 Regulatory recognition elements and RNA binding

Post-transcriptional gene-expression regulation generally involves a direct interaction between a *trans* factor and the regulatory recognition element (RRE) on the RNA transcript being regulated. Figure 2.8 illustrates four types of interactions with RNA: RNA-RNA interaction, protein-RNA interaction, a small-molecular ligand or a peptide ligand binding to RNA. The first two involve ncRNAs or RBPs that interact specifically with an RRE—a local RNA sequence and/or structure motif. In contrast, molecular or peptide ligands bind to a (tertiary) structural pocket of RNA and this usually changes or stabilises the local RNA structure, as is the case for many riboswitches [29]. Regulatory RNA-RNA interactions usually require that the RRE is accessible (i.e. unpaired). RBPs can bind specifically to both sequence and structure.

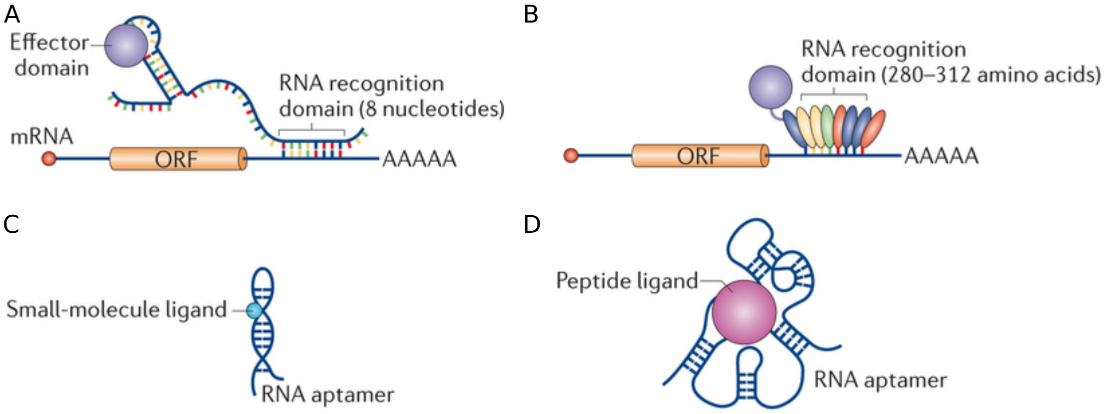


Figure 2.8. Regulatory recognition elements on RNA. An illustration of four types of interaction with RNA regulatory recognition elements (RREs): (A) RNA-RNA interactions, usually with *trans*-encoded ncRNA; (B) protein-RNA interactions with local RNA motifs; (C) small molecular ligands or (D) peptide ligands binding to tertiary structure pockets where such interactions usually change or stabilise local structure. Illustration adapted with permission from Macmillan Publishers Ltd: *Nature Reviews Molecular Cell Biology* [101], copyright 2013 (license no. 3363740908111).

An RNA-recognition protein domain requires about 40 amino acids to specifically recognise a single nucleotide. Hence, eight nucleotides are recognised by a domain of about 320 amino acids, encoded by 960 nucleotides. In contrast to proteins, most RNA-guided regulation requires a short *seed* interaction involving about 6–8 consecutive base pairs. This seed interaction can be extended further by additional base pairs, but most RNA-RNA interactions do not exceed 25nt [257]. This large difference means that regulation based on RNA-RNA interactions is more efficient, however, proteins remain essential in regulatory processes and protein-RNA interactions are required as an interface for the formation of large RNA-nucleoprotein complexes that perform elaborate functions [154, 156, 245, 295]. In this dissertation, there is a greater focus on RNA-RNA interactions than on protein-RNA interactions; although the latter often coincides with the former to stabilise RNA-RNA interactions.

2.7.1 Experimental detection of interactions with RNA

The greatest resolution of both RNA-RNA and RNA-RBP interactions is given by methods that derive the three-dimensional configuration of the molecular complex, such as nuclear magnetic resonance or x-ray crystallography (see e.g. [326, 330]). A further approach to detect binding sites of both ncRNA and RBPs is to perform mutational experiments where both the sequence and structure (when required) of the RRE are mutated. The effect of every mutation is compared with the wild type. Decreased or abolished signal[1] is indicative of an interaction at the mutated site. For example, mutations were performed to detect CRISPR processing by Cas6 in CRISPR-Cas systems [227, 298]. Due to time and cost expense of the above approaches, they can only be applied to individual cases with known or hypothesised RRE sites and cannot be used to detect novel RREs on a transcriptome-wide scale.

Changing the expression of an RBP or ncRNA can infer their binding partners, but not sites of RREs. For example, in the search for miRNA targets, either a specific miRNA was overexpressed, or the dicer protein, which is essential for mature miRNA processing, is knocked out: the affect on transcript or protein expression levels is measured and significantly increased or decreased expression levels are considered to be evidence of miRNA regulation [276,279,285]. Changing the expression of a single gene, however, can affect many dependent processes; therefore, not only direct effects are measured. Moreover, results are skewed due to the fact that highly abundant molecules are more likely to find a binding partner than rare molecules.

Immunoprecipitation-based protocols have been developed for the detection of RREs bound by a specific RBP. Immunoprecipitation (IP) requires a strong affinity between RBP and RRE, and only whole (or longer segments of) target transcripts are identified when only a simple IP method is applied, e.g., the `RIP` approach [243]. Therefore, protocols were developed to stabilise interactions by forming covalent bonds between RBP and the RRE. The bonds are induced by UV light or formaldehyde in a process called *cross linking* [316]. RNA sequence not protected by the cross-linked RBP can be digested to gain more specific information on the RRE location. After sequencing the bound RNA, the `RNA-seq` data is mapped to the genome and profiles are compared with a control to detect significantly enriched peaks of

[1] The signal that is triggered by an interaction with the RRE, e.g., the observation of an expected phenotype, a change in expression level(s), or of a specially designed reporter system.

overlapping reads that mark potential RRE sites. This general approach is termed CLIP-seq and variants include PAR-CLIP [9], HITS-CLIP [58,352] and iCLIP [147,174,300]. Interactions between ncRNAs and their target RREs are usually facilitated by an RBP (e.g., AGO in miRNA binding, Section 2.2.1), therefore, CLIP-seq experiments can also be applied to the detection of ncRNA binding. In the case of miRNAs, target sites are identified by CLIP-seq of one or all of the AGO proteins [46,122,355]. However, this data does not provide the miRNA involved in the interaction. For this purpose, a variant protocol (CLASH) was established that ligates the miRNA sequence with the RRE [135].

CLIP-seq is considered the gold standard for detecting RBP and miRNA targets but despite its great success there are still caveats: (1) the data may contain many false positives due to inherent noise [55,318]; (2) a large number of binding sites remain unidentified (a high false-negative rate) because CLIP-seq is sensitive to expression levels and is both time and tissue dependent [24]; (3) limited mappability [61] and mapping difficulties at splice sites lead to further false negatives, even on highly expressed mRNAs. Consequently, CLIP-seq experiments should be complemented with the computational discovery of missing binding sites.

2.7.2 RNA-RNA-interaction prediction

An ncRNA regulates its target RNA by physically associating with its RRE to form an intermolecular duplex (Figure 2.8), which follows similar rules to intramolecular RNA structure formation (Section 2.4). The *complementarity* between ncRNA and RRE is a central feature in most approaches for predicting RNA-RNA regulatory interactions. Existing prediction methods differ in their means for measuring the degree of complementarity.

The duplex formation involves the bonding of complementary base pairs between the ncRNA and the RRE. Hence, the degree of complementarity can be measured by an extended secondary structure prediction model that determines the thermodynamic stability of the RNA-RNA duplex. One of the first approaches in this direction was RNAhybrid[1] [252]. It uses the usual energy model for secondary structures [210], but considers only stacking and internal loops in duplex formations. This simple model, however, disregards internal mRNA structure that might block an RRE site. More advanced models exist that also include intramolecular base pairing [35,47,146] but are too slow for predicting ncRNA targets on a genome-wide scale.

For very short ncRNA, such as miRNA and siRNA, it is unlikely that an internal structure affects binding efficiency due to their integration into an AGO protein [78]. For longer RNA, it is possible to use a simplified measure of mRNA intramolecular structure, called accessibility (see Section 2.5.6). An RRE is *accessible* if it is not involved in internal mRNA base pairs and is thus free for the interaction with the ncRNA. RNAup [224] defined accessibility as the probability that a specific region of the mRNA (in our case the RRE) is single-stranded in the ensemble of all possible structures and this measure has been shown to be significant for

[1] (also implemented in RNAduplex, *Vienna RNA Package* [119]), or later in targetRNA [314]

successful binding to RREs [164, 258]. Again, the computational complexity of RNAup was too high for genome-wide scans, and thus two solutions exist with more practical run-times: (1) IntaRNA [34] uses a full energy calculation for the accessibility, but a heuristic method (including a seed condition) for determining the best duplex structure; (2) RNAplex [304] uses a position-specific penalty score that depends only on a local context to approximate the accessibility of a complete region.

In bacteria, RNA-RNA interaction is a predominant form of gene regulation [193,325] involving a class of small ncRNAs termed sRNA. Although IntaRNA was specifically developed for predicting such interactions [34, 259], it still produces many false positive predictions [258]. Many types of sRNA are highly conserved across many prokaryote species, thus the use of conservation can greatly reduce the number of false positive predictions. PETcofold [282,283] uses interaction site sequence conservation, which is a very restrictive method. The better use of conservation information is to detect pairs of sRNA–targets that are conserved across many species, as was developed recently in CopraRNA [343]. Using conservation to filter miRNA target predictions leads to a comparatively low sensitivity [250, 251] because miRNA can target up to hundreds of genes and many target sites are not conserved. However, the CopraRNA approach has not yet been implemented for miRNA but is expected to lead to better results in the future.

2.7.3 MicroRNA target prediction

A mature miRNA is incorporated into an AGO protein and guides the entire RNA-induced silencing complex (RISC) to its target at an RRE specific to miRNA, an MRE (miRNA recognition element). Most miRNA target prediction tools base their initial search on thermodynamic stability between miRNA and MRE. For plant miRNAs, the application of thermodynamic stability is almost sufficient for predicting putative MREs. To this end RNAhybrid has been applied successfully to identifying miRNA targets in the model plant organism, *Arabidopsis thaliana* [178,252]. In comparison with plants, animal duplex structures are much less stable, therefore, assessing the degree of complementarity via duplex stability alone leads to a vast number of candidate MRE sites and does not perform well as a sole predictor [323].

As for any prediction approach for RNA-RNA interactions, the use of additional features can increase the specificity of target prediction. A key determinant of miRNA target specificity is the well-defined seed interaction of six uninterrupted base pairs between nucleotides 2–8 of the miRNA and the MRE, and various extensions and definitions of this region [11, 166, 190, 191, 285]. Further determinant features are conservation, MRE context information, special characteristics of the duplex formation (e.g. compensatory 3' binding), overrepresented seed motifs, multiple MREs per target, cooperative RBP binding, relative MRE position, accessibility and AU content of the MRE and the direct sequence context, and expression levels of both miRNA and mRNA. An evaluation of the relative contributions of these features for detecting miRNA targets were published in [132, 323]. Most tools developed specifically for miRNA target prediction use a unique combination of these filter features to achieve a higher specificity. Some popular and/or more recently developed examples are given by PITA [164], Pictar [183], MiRanda [79, 153], TargetScan [92], rna22 [216], EIMMo [93], miRmap [323], MREdictor [148]. More recently explored features are the expression levels of miRNAs and target mRNAs in tissue-specific cells [225,246,262]; extended context information, e.g., sequence composition, length, and structure [113,132,171]; MRE sites in coding sequences [94,122,250,277], and cooperative or competing factors of multiple MRE sites per target mRNA [113,205,271,272] and protein-binding sites, e.g., AU-rich elements [21, 77,150]. Reviews of prediction performances are given in [3,148,204,205,250,251,323]. In these comparisons it is clear that a simple seed requirement results in the most sensitive predictions and that the use of conservation greatly increases the precision of predictions.

Most tools use a subset of features to achieve target predictions, however, to model the full spectrum, machine learning techniques are required to select the most informative features and to avoid overfitting of models to the data. These techniques are more suited to modelling the complex interplay between features used and the problem of learning which features (or combination of features) contain the most information with respect to the prediction performance. A large number of such approaches have been developed recently, a subset of notable examples include mirSVR [20], Targetminer [13], MTar [41], MultiMiTar [217] DIANA-microT-CDS [250] and TargetSpy [299]. Machine learning approaches mainly differ in which kind of features and machine learning techniques are used to score the predictions and in the quality of the training and testing data. A variety of machine learning techniques have been applied to miRNA target prediction, although support vector machines (SVMs) are a preferred method. SVMs are used by Targetminer and MultiMiTar. The use of support vector regression is able to predict the strength of miRNA-MRE interaction (as in mirSVR), however, this requires training data on the strength of regulation.

The selection of appropriate positive and negative data is an important, but challenging task. With respect to the positive data, mirSVM uses MiRanda predicted targets, and other approaches usually use experimental data. These are collections of either experimentally verified miRNA-MRE interaction sites (used by Targetminer, MTar, MultiMiTar), or from high-throughput experiments. Besides array-based approaches [113], one common source is CLIP-seq [46, 122], which provides genome-wide information on the binding site of the AGO protein, used by DIANA-microT-CDS, mirSVR. In most CLIP-seq protocols the miRNA is not measured and has to be inferred from the data. Some tools have emerged to make predictions on miRNA binding partners for CLIP-seq data [49,166,347,348]. Furthermore, pSILAC is a high-throughput method to directly measure changes in protein synthesis [285] and to overcome the problem that for a verified MRE, the impact on protein level remains unknown (used by MultiMiTar and DIANA-microT-CDS). TargetSpy used the pSILAC data for validation instead of training.

Early approaches generated negative data by selecting randomly generated sequences, however, this is not a good choice since they are too distant from real negative examples, i.e., false positive predictions. Hence, it is important to carefully generate a negative data set. This is even more problematic than the positive examples, since no “gold standard” for negative examples exists. TargetMiner, for example, generated an accurate set of negative examples from a pool of predicted but experimentally not validated target interactions and Yousef and colleagues implemented a one-class technique [350].

Extended reviews of computational miRNA target prediction are given in [213, 215, 305, 310, 342].

Part II

Conservation of regulatory motifs

Part II: Conservation of regulatory motifs

Many roads. One goal. All roads lead to Rome.—Alain de Lille and Geoffrey Chaucer

Conservation across many species can provide a powerful tool for detecting regulatory function of non-coding RNA (ncRNA): generally speaking, the more distant the species that harbour a common signal, the greater the evidence of evolutionary pressure to conserve the signal. In this part, we explore the conservation of CRISPRs to detect and characterise binding motifs for Cas proteins. First, a clustering procedure for characterising all available CRISPRs is presented in Chapter 3 and application scenarios are highlighted in Chapter 4. Overall, this part exemplifies the use of conservation to characterise regulatory recognition element (RRE) motifs. In this case, we know the approximate location of the RRE because the Cas protein binds to the repeat of the CRISPR RNA. Hence, characterising the RRE is easier than when its location is unknown. It would be possible to apply a procedure similar to the one proposed here to characterise RREs that have either been experimentally verified or predicted with high certainties. However, the techniques used here were tailored to the specific requirements of the CRISPR-Cas system. The work presented in Chapters 3 and 4 were a part of the following publications: [P1–P3, P5, P9, P11].

CHAPTER 3

CRISPRmap: Repeat conservation in CRISPR-Cas systems

Despite the conceptual simplicity of the underlying mechanism, a large variety of distinct CRISPR-Cas systems exist (see Section 2.2.2). This variety leads to the necessity of categorising systems into groups for which members of a single group are functionally related. The characterisation of CRISPR-Cas systems helps to make assumptions across related systems. CRISPRs are associated with distinct sets of Cas proteins. In the literature, a CRISPR-Cas system is usually characterised by the encoded Cas proteins into at least 10 widespread subtype annotations [123,201,202,324]. Although the Cas-centric classifications of CRISPR-Cas systems is generally effective, an accurate Cas-protein–based classification is complicated: Many of the *cas* genes belong to extremely diverse families [123,202]; CRISPR loci may include novel, chimeric, mixed subtypes, or *cas* genes that are missing entirely [98,155,202,260,287]; and it is not always obvious which *cas* genes are specific to a repeat-spacer array or Cas proteins could be shared between arrays [260].

In this chapter, we present a comprehensive classification of all publicly available CRISPRs that is based solely on the sequence and structure evolution of repeats. The repeat-spacer array is the only element present in all CRISPR-Cas systems. Therefore, these systems are identified first by the existence of such an array. In contrast to the annotation of *cas* genes, repeat-spacer arrays are easily identified by programs such as `CRISPRFinder` [115] or `CRT` [23]. The repeat is the central regulatory element in the CRISPR-Cas system as it serves as a binding template for Cas proteins in all three phases of immunity: adaptation, interference and crRNA maturation (Section 2.2.2). For these reasons, a systematic repeat-based classification is fundamental for extending knowledge about the function, diversity, and phylogeny of CRISPR-Cas immune systems. A phylogenetic study of these immune systems is not trivial because entire (or elements of) CRISPR-Cas systems are frequently transferred between unrelated species. Thus, their evolution does not always follow the evolution of the host genomes [144,202].

Similarities between CRISPRs are assumed to reflect conserved binding motifs and mechanisms. The binding affinity of Cas proteins is not only affected by the repeat sequence: a small hairpin structure is a key binding motif for Cas endoribonucleases in several systems [32, 103, 129, 131, 227, 274, 298] [P7, P10]. To correctly identify these structure motifs, our clustering is the first that is based not only on sequence—but also on structure—similarities. This approach is well-established for the identification and characterisation of structured ncRNA [134, 240, 338, 339] (*c.f.* Section 2.6). For these ncRNAs, the conservation of structure is often more important than sequence for the biological function [95, 117]. Although CRISPRs are partially structured ncRNAs, no structure-based clustering exists. To our knowledge, the only CRISPR-specific classification was performed on 349 bacterial and archaeal repeats in 2007 [180]. Although structure motifs were identified, the underlying clustering was based purely on sequence and not structure similarity. An analysis of the archaeal domain, also based on only sequence similarities, was done more recently [98].

Since at least a third of CRISPRs do not contain structure motifs, we performed an independent clustering of CRISPRs based solely on sequence similarities to identify conserved sequence families. Independent sequence-and-structure and sequence-only clusters provide a more complete overview of the conservation of both unstructured and structured CRISPRs. We combined identified structure motifs and sequence families with a hierarchical representation of sequence and structure similarities to generate a map that directly reflects relationships between classes and individual CRISPRs. This hierarchical `CRISPRmap` tree enables a fast comparison between CRISPRs of interest and previously published systems. Automated access to our data via an easy-to-use web server (`CRISPRmap`[1]) allows users to identify relative positions of both published and unpublished sequences. `CRISPRmap` is a valuable resource to elucidate and generalise functional mechanisms of CRISPR-Cas immunity. This chapter was adapted from [P3].

3.1 CRISPR-Cas data collection and annotation

In addition to generating a comprehensive set of CRISPRs, we derived automated processes for annotating *cas* genes and Cas subtypes; and mapped these to a CRISPR locus. Finally, we generated Cas1 clusters to determine the link between Cas protein and CRISPR RNA evolution.

3.1.1 CRISPR data

All available genome sequences were downloaded from the NCBI server (http://www.ncbi.nlm.nih.gov/) and the CRISPR databases: CRISPI [269] and CRISPRdb [114] (August 2012). Redundant genomes were removed. We predicted CRISPRs using the two most commonly used programs, `CRISPRFinder` [115] and `CRT` [23]. For both tools, we used parameters that corresponded to at least three repeats within an array and the repeat and spacer lengths

[1] The results presented in this chapter are from `CRISPRmap` version 1.0.

Table 3.1. Summary of our `REPEATS` dataset including all publicly available CRISPR arrays.

	Archaea		Bacteria	
Genomes	279		2,289	
Genomes with CRISPRs (percent)	177	(63 %)	877	(38 %)
Plasmids	41		1,286	
Plasmids with CRISPRs (percent)	14	(34 %)	76	(6 %)
CRISPRs	643		2,884	
Repeats per array (median)	3–190	(15)	3–1371	(12)
Repeat lengths (median)	20–44	(29)	19–48	(30)
Spacer lengths (median)	20–50	(38)	19–70	(35)

were set to 18–58 nt. Although repeats within one array are largely identical, they can contain some mutations, especially towards the 3' end of the array. Thus, we used a single representative repeat of a CRISPR array by calculating the consensus sequence of all repeat occurrences. Finally, we merged the results from both programs and the CRISPR databases to form a non-redundant set, which we refer to as `REPEATS`. Table 3.1 gives a summary of our `REPEATS` dataset.

Using the described procedure, we obtained over 3,500 consensus repeat sequences from predicted CRISPR arrays in ∼2,500 available genomes. 63 % of archaea and 38 % of bacteria contained predicted CRISPR arrays, similar to previous observations [114, 144, 202]. Interestingly, the number of plasmids that contained CRISPR arrays is considerably lower: 34 % and 6 % in archaea and bacteria, respectively. Thus, most CRISPR arrays (94 %) are located on chromosomes. This dataset is the most complete set of CRISPRs to date; we compared the `REPEATS` dataset to previous work in Figure D.5.

The results from `CRISPRFinder` and `CRT` give no information on the correct strand orientation. Therefore, we predict the repeat orientation within our clustering approach. To do this we required CRISPR data with known orientations. The following two sets were gathered for this purpose:

- **Set of repeats from Kunin *et al.* 2007.** We downloaded the dataset from the supplementary material of [180] and refer to it as $\texttt{REPEATS}_{Kunin}$. This dataset contains 271 bacterial and 78 archaeal sequences (349 in total). The orientations were predicted by the authors using previously published sequence features.

- **Set of archaeal repeats from Shah and Garrett 2011.** We received 378 archaeal repeat sequences from Shah and Garrett that were the basis for the results in [287]. The repeat orientations were manually verified by Shah and Garrett. We refer to this dataset as $\texttt{REPEATS}_{Shah}$.

3.1.2 Cas gene and Cas-subtype annotations

Annotations of all *cas* genes. Subtype independent annotation of *cas* genes was performed on the entire chromosome or plasmid which harbours the respective CRISPR array.

We applied the TIGRFAM models from Haft *et al.* [123, 124] in combination with HMMER [75] but used the more recent *cas* gene names from Makarova *et al.* [202]. A *cas* gene was annotated when one of its respective models was found with an E-value ≤ 0.001. On our web server site, we provide a full table of *cas* gene annotations for each repeat, giving the minimum distance of that gene to the CRISPR array. For each sequence family and structure motif, we identified single *cas* genes that were associated with the majority of CRISPRs in the respective class; all *cas* genes on the entire chromosome or plasmid with the CRISPR were considered. Results are given in summary in Tables D.2–D.19.

Cas subtype annotation from Makarova *et al.* 2011. The automatic annotation of subtypes is tricky due to the fact that genes of multiple subtypes can be present in the genome, subtypes are often incomplete, and it is not known if the *cas* genes must be within a certain distance of the CRISPR array. However, in many published CRISPR-Cas systems, the *cas* genes are located either directly upstream or downstream of the array [202]. We used the following procedure that enabled a suitable trade-off between precision and recall of the annotations: We first compiled a list of signature *cas* genes that were unique to each type and subtype from [202][1]. For each repeat, i.e., CRISPR array locus, we identified first the closest subtype signature and then noted the distance of the respective type signature, if available. We plotted the distance of subtype and type signatures and determined a clear peak (at 14.5 kb) in their distances to their respective CRISPR array (Figure D.1). We considered a cutoff of 180 kb to represent a suitable distance from the CRISPR array; this cutoff corresponds to the 70th percentile of distances of the subtype signatures. A repeat is assigned to a subtype if both subtype and type signatures are within this distance. Note that with this approach, not all *cas* genes have to be present or annotated.

Clustering of Cas1 proteins. Cas1 protein sequences were assigned to the closest CRISPRs if they were within 180 kb of the array (see Figure D.1 for cutoff explanation). These Cas1 proteins were clustered using Markov clustering (`MCL`) [80, 320] with default parameters. The `MCL` method is a popular method for clustering biological sequence data and was applied previously to CRISPRs [180, 287]. Here, pairwise protein-sequence similarities were calculated with the local Smith-Waterman alignment algorithm [292] and percent protein identities below 40 % were set to zero to reduce noise. Only clusters with at least ten proteins were considered.

3.2 Detecting sequence and structure conservation independently

We performed a comprehensive search for both conserved sequence families and small CRISPR-like hairpin motifs, using *independent* approaches to allow for both structured

[1] Please note that the recently updated classification presented in [324] was published after this work was performed and is not considered here.

and unstructured repeats. First, we partitioned CRISPRs into sequence families using MCL, as in previous studies [180, 287]; in addition, we applied sequence profiles to refine the MCL clusters (Section 3.2.1). With this procedure, we identified 40 conserved families. The mean pairwise nucleotide-sequence identity of 82 % (68–96 % for each family) reflects a high level of sequence conservation. Second, independent to identified sequence families, we searched for conserved structure motifs using sequence-and-structure alignments (Section 3.2.2). Structure motif candidates were constrained to be reminiscent of those previously published [32, 129, 227, 274, 298] [P7, P10]. More specifically, 33 small hairpin (or stem-loop) motifs with at least four base pairs and no bulges were identified. Their sequence conservation was generally lower than for sequence families: mean pairwise sequence identities ranged between 47–94 % with an average of 69 %. Sequence families and structure motifs were numbered according to cluster size, starting with the largest clusters; the smallest cluster contained 10 sequences. Summary tables with sequence logos for families, secondary structures for motifs, mappings between families and motifs, and annotations are available in Section D.1; full alignments are available on the CRISPRmap web server, version 1.0.

To provide further support for our secondary structure predictions, we evaluated the motifs using the general ncRNA predictor, RNAz [118]. Although RNAz is not specifically trained for CRISPR elements, it classified 79 % (26 out of 33) of our motifs as structured ncRNAs with an SVM-RNA-class probability greater than 0.6 (22 motifs even achieved over 0.9; a clear indication that these motifs are evolutionary conserved). Compared to other ncRNA classes, RNAz only exhibits such promising sensitivities on some of the classical ncRNAs [266, 267], for example, transfer RNAs or miRNAs, which are known for their distinct and well-defined secondary structures [105, 157].

In total, out of all CRISPRs in our REPEATS dataset, 64 % were assigned to a conserved sequence family and 51 % were assigned to a structure motif. 26 % of repeats remained unassigned to either a family or motif, i.e., showed no conservation with available CRISPRs.

3.2.1 Clustering of repeat sequences into conserved sequence families

Repeat sequences were clustered into related families based on global sequence similarity using MCL [80, 320] (downloaded from http://micans.org/mcl/). First, we calculated pairwise similarities with the Needleman-Wunsch alignment algorithm [230]. These nucleotide-sequence similarities (i.e., percent identities) were plotted (Figure D.2) and a reasonable cutoff of 65 % nucleotide identity was chosen to represent sufficient similarity. Similarities below this value were explicitly set to zero to reduce noise. We ran the MCL program with an inflation parameter $I = 2.5$. This parameter gave a good balance between the number of sequences assigned to a family and the conservation within a family. Only clusters with at least ten repeat sequences were considered as a *conserved* sequence family.

We supplemented the MCL clustering with sequence profiles generated by CLUSTAL W [311], version 1.83. We used these profiles to re-assign repeats to families to which they were sufficiently similar, as follows: Let $sim(F, r)$ be the profile score of a repeat r compared with the profile of the family F, where $r \notin F$. For each original family, the minimum (F_{min}) and maximum (F_{max}) profile similarity was determined by removing each sequence from the family, re-calculating the profile for the remaining sequences, and determining the similarity score of the respective repeat to the profile. A repeat r was then assigned to a sequence family F if $sim(F, r) \geq F_{min}$ and the distance between $sim(F, r)$ and F_{max} is the minimum compared to all other families. In total, 73 sequences were re-assigned by the sequence profiles. The sequence conservation did not change significantly, but we were able to identify those few repeats that where missed by the `MCL` algorithm.

For each family, we generated sequence logos (Tables D.2–D.19) using a multiple sequence alignment computed with `MAFFT` [159], version 6.4. The multiple sequence alignment was converted into a logo by `WebLogo` version 3 [57].

3.2.2 Identifying conserved structure motifs

Our procedure for identifying conserved, local, hairpin-structure motifs (referred to as structure motifs) in all CRISPRs involves a complex, multi-faceted workflow.

Step 1—pool of repeats. The procedure starts with a pool, P_u, of repeats that have not been assigned to a structure motif. Initially P_u contains our entire `REPEATS` dataset. The orientation of each repeat is predicted by a graph-kernel-based machine learning model [56], slightly modified to work on directed graphs. We trained the model on the $\texttt{REPEATS}_{Shah}$ dataset (using the 253 repeats that had less than 95 % similarity to ones in $\texttt{REPEATS}_{Kunin}$). Each repeat sequence is given as a directed graph, i.e., the nucleotides are represented by nodes. These are linked by directed edges indicating the particular orientation. To test the performance of our model, we applied it to the $\texttt{REPEATS}_{Kunin}$ dataset. Overall, we achieved a performance of 0.68 AUROC when using the feature parameters radius $r = 1$ and distance $D = 2$. Since we did not achieve a perfect orientation prediction (mostly due to insufficient training data), we addressed this issue throughout our clustering process. Nonetheless, the model ensures that at least the majority of sequences are in the correct orientation for the first clustering steps.

Step 2—generating a hierarchical cluster tree reflecting sequence and structure similarity. A hierarchical cluster tree T_i for the current iteration i is generated from all sequences in P_u using `RNAclust` [339]. `RNAclust` employs a hierarchical clustering algorithm (UPGMA [116]) based on similarities calculated with a sequence-and-structure alignment program, `LocARNA` [338, 339]. Thus, relationships in the resulting binary tree not only reflect sequence, but also structure similarity. For each node of the cluster tree, there exists a sequence-structure alignment with the respective predicted consensus structure as given by `LocARNA`.

Step 3—selecting subtrees with CRISPR-like consensus structures. Starting from the root node in T_i, each child node is traversed in hierarchical order until a CRISPR-like hairpin consensus structure is found at a certain node t. The consensus structure is *local* in the sense that it does not cover the entire repeat sequence. All repeats descending from node t are considered to form a candidate structure motif, $\mathit{Motif}(t, T_i)$, if the following requirements, derived from published repeat structures [32,103,129,131,227,274,298], are met:

1. The consensus structure of $\mathit{Motif}(t, T_i)$ is a hairpin with a stack of at least four base pairs and no bulges or internal loops.
2. At least 10 repeat sequences fit to the consensus structure of the motif candidate; repeats that do not fit to the consensus structure are removed from $\mathit{Motif}(t, T_i)$.
3. The two direct child nodes of t must have compatible consensus structures, which we define as having ≥75 % of the base pairs overlap with the consensus structure at t.

If the requirements for $\mathit{Motif}(t, T_i)$ are met, then all descendent nodes of t are assigned to $\mathit{Motif}(t, T_i)$ and the procedure is repeated until all nodes in T_i have been checked for belonging to a structure motif.

Step 4—supertree of only structured repeats. All repeats that have not been assigned to a structure motif are removed from the tree and are put back into the pool of unassigned repeats P_u. All other repeats, which form one of the consensus structures, are put into a set P_s. From this set P_s a *supertree*, $ST(i)$, is generated by repeating Steps 2 and 3. Again repeats that do not conform to the criteria are removed and put back into the unassigned pool P_u. This re-clustering ensures the robustness of identified motifs.

Step 5—merging supertrees. In one `RNAclust` run, we identify conserved structures of repeat sequences that are neighbouring in the cluster tree T_i. To locate more distantly related repeat sequences that can still form a common consensus structure, we repeat the clustering with the remaining sequences in the pool P_u. Consequently, Steps 2–4 are repeated for three iterations, resulting in three separate supertrees (ST_1, ST_2, and ST_3) that are merged into one supertree, $ST_{1,2,3}$. Merging starts with ST_1: Since it is the result of the first iteration, it includes the largest and most well-conserved structure motifs. Each structure motif of the supertrees ST_2 and ST_3 is merged with ST_1, one at a time. Due to the orientation uncertainty, we also attempt to merge the reverse complement sequences of the whole structure motif. Merging occurs by repeating Steps 2–4 and we use the orientation that results in the fewest number of repeat sequences being lost to P_u in the merging process.

Step 6—final cluster tree with structure motifs. We perform a last post-processing step to produce the final cluster tree with the structure motifs. For each structure motif, we calculate the consensus structure of the reverse complement repeat sequences. *GU* base pairs become A and C bases and cannot pair in the reverse complement orientation. Therefore, we consider the orientation with the most stable consensus structure to be correct. We also check whether the reverse complement of a motif can be merged with another existing motif. Two features are common to CRISPR sequences: a conserved 3' end of repeats, $AUUGAAAC/C$ and a majority of A instead of U nucleotides for archaeal sequences—as observed in the manually verified orientations in $\texttt{REPEATS}_{Shah}$. We checked the consensus sequence of all CRISPRs belonging to a motif in both possible orientations for the existence of one of the above features. This information was used to derive the correct orientation of a motif. If any changes were made in the original orientation, the orientations of the respective CRISPRs were swapped and Steps 2–4 were repeated for all CRISPRs currently assigned to a motif. Note that changes to the input set can lead to changes in the resulting tree, therefore, our repeated runs of `RNAclust` ensure that most of the noise is removed and we only include stable structure motifs in our final result.

Improving the orientation of repeats in our `REPEATS` data. The identification of conserved structure motifs gives some evidence on the likely orientation of the repeats involved. For repeats not assigned to structure motifs, however, we had no information to deduce the correct orientation. Therefore, we merged all structured repeats with the original $\texttt{REPEATS}_{Shah}$ data and re-trained our prediction model; we excluded repeats ≥ 95 % similarity with the test data. By doing this, we assume that the majority structured repeats have correct orienations after our clustering procedure and thus we can extend the original set of repeats with "known" orientations. Again, we tested our model on the $\texttt{REPEATS}_{Kunin}$ data and achieved a substantial improvement with an AUROC of 0.82 in comparison with 0.68 previously. We subsequently used our re-trained model to predict the correct orientation of the repeats remaining in the unassigned pool P_u. Even if some orientations are still incorrect, this step ensures that the repeat orientations in our `REPEATS` data are consistent. To add the sequences that were previously in the incorrect orientation, we repeated Steps 1–6 with the improved orientation predictions.

3.3 A visual map of CRISPR conservation

As a visual *map* of both bacterial and archaeal CRISPR domains, we combined our discrete categorisation into conserved families and motifs with a hierarchical tree, based on sequence-and-structure similarities (compared with a non-hierarchical, sequence-similarity-based visualisation in Figure D.8). The so-called `CRISPRmap` tree was generated by `RNAclust` [339] and visualised with `iTOL` [189]. The tree reflects relationships based on sequence *and* structure similarity; however, when a repeat is unstructured, only the sequence similarity is considered. At a single glance, the `CRISPRmap` tree details relationships between individual repeats and whole families and motifs (Figure 3.1).

In addition to the repeat families and motifs, we annotated taxonomic phyla, Cas1 sequence homology clusters, and Cas subtype annotations [123, 202]. Each leaf represents CRISPR sequence and the leaf branches are coloured according to whether the CRISPR stems from bacteria or archaea. Figure 3.1 shows one possible view of the CRISPRmap tree with sequence-families, structure-motifs, superclass classifications (see Section 3.3.1) and the domain. Further views and annotation data are available in the supplementary material and on our CRISPRmap web server: http://rna.informatik.uni-freiburg.de/CRISPRmap.

In summary, the CRISPRmap tree was designed to provide a visual overview of CRISPR conservation and to aid in the understanding of CRISPR-Cas diversity.

3.3.1 The CRISPRmap tree is divided into six superclasses

Based on sequence-and-structure similarities and the tree topology, the REPEATS dataset could be broadly grouped into six major superclasses (Figure 3.2). The superclasses, labelled A–F, are ordered according to generally decreasing conservation. The following information is quickly observed in the CRISPRmap tree (Figure 3.1): Superclass A contains highly conserved CRISPRs on the sequence level, but only a few structure motifs without many CRISPRs assigned to them. Superclasses B–C contain sequence families that roughly correspond to one structure motif each; the same is true for half of superclass D. The other half of superclass D and superclass E contain very little sequence conservation, but many conserved motifs containing fewer CRISPRs. Archaeal CRISPRs in both superclasses A and F contain well-conserved sequence families and we find structure motifs for about half, however, these are less stable than the bacterial motifs in superclasses B–D (Tables D.2–D.19). The bacterial repeats in superclass F are very divergent: We included arrays with at least three repeat instances to ensure that our dataset was complete. Many arrays with up to five repeat instances, however, show little conservation (Figure D.7): roughly 50 % were not assigned to sequence families or structure motifs and most are in this diverse part of superclass F. In addition to array size, we marked repeats or (average) spacers with unusual lengths on the CRISPRmap tree in Figure D.7. Some of the very short arrays, especially those with unusual repeat and/or spacer lengths are unlikely to contain functional CRISPRs.

We summarised annotations and clustering results to give a brief overview of each superclass in Figure 3.2; more details are given in the following results. In the CRISPRmap tree views (e.g., Figure 3.1), the superclass is always annotated in the outer-most ring. Note that missing data points (i.e. repeats) in the CRISPRmap tree induces noise in the tree topology. Therefore, increasing the number of repeats in the CRISPRmap database will most likely increase the accuracy of the tree.

3.4 An in-depth analysis of clustering results

We analysed the CRISPRmap data to gain further biological insights into aspects of CRISPR-Cas systems. In detail, we looked at motifs at cleavage sites, variations in conservation patterns, the link between CRISPR and Cas subtype evolution, and the evolution or transfer of CRISPR-Cas systems among different species.

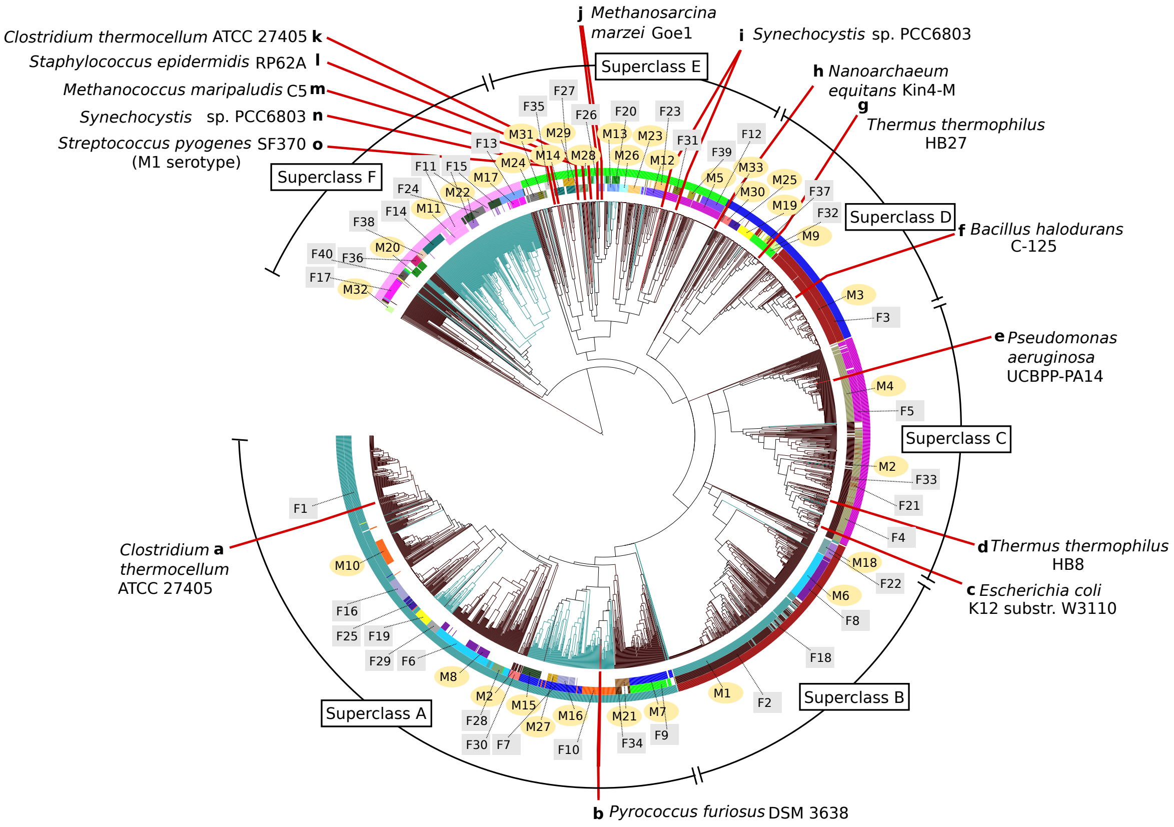

j Methanosarcina marzei Goe1
i Synechocystis sp. PCC6803
h Nanoarchaeum equitans Kin4-M
g Thermus thermophilus HB27
f Bacillus halodurans C-125
e Pseudomonas aeruginosa UCBPP-PA14
d Thermus thermophilus HB8
c Escherichia coli K12 substr. W3110
b Pyrococcus furiosus DSM 3638
Clostridium thermocellum ATCC 27405 a
Clostridium thermocellum ATCC 27405 k
Staphylococcus epidermidis RP62A l
Methanococcus maripaludis C5 m
Synechocystis sp. PCC6803 n
Streptococcus pyogenes SF370 o
(M1 serotype)
Superclass A
Superclass B
Superclass C
Superclass D
Superclass E
Superclass F
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
F11
F12
F13
F14
F15
F16
F17
F18
F19
F20
F21
F22
F23
F24
F25
F26
F27
F28
F29
F30
F31
F32
F33
F34
F35
F36
F37
F38
F39
F40
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
M21
M22
M23
M24
M25
M26
M27
M28
M29
M30
M31
M32
M33

Figure 3.1. The CRISPRmap tree: a map of repeat sequence and structure conservation. The hierarchical tree is generated with respect to repeat sequence and structure pairwise similarity and the branches are coloured according to their occurrence in the domains bacteria (dark brown) or archaea (blue-green). The rings annotate the conserved structure motifs (inner), sequence families (middle), and the superclass (outer). Motifs and families are marked and highlighted with yellow circles, and grey squares, respectively. Finally, we marked locations (a–o) of published CRISPR-Cas systems for which experimental evidence of the processing mechanism exists [32, 60, 99, 103, 129–131, 227, 248, 260, 274, 298, 326] [P7, P10]. A summary for these published systems is given in Table 3.2. Repeats that show no conservation, i.e., were not assigned to either a sequence family or structure motif, were removed to clarify the visualisation. Figure taken from [P3].

Table 3.2. Published CRISPR-Cas systems with experimental evidence of the processing mechanism. In particular, these are systems for which the Cas endoribonuclease is characterised and/or the repeat structure has been verified. Published results are consistent with our data. The IDs, a–o, are marked, in order, as red lines on the CRISPRmap tree in the manuscript in Figure 3.1. Table taken from [P3].

ID	Organism	Family	Motif	Cas Subtype	Summary
Superclass A					
a	*Clostridium thermocellum* ATCC 27405	F1	-	I-B	Unstructured; 8-nt 5' tag; biochemical evidence to show **Cas6b** activity [260]
b	*Pyrococcus furiosus* DSM 3638	F10	-	III-B	Unstructured; 8-nt 5' tag; cleavage by **Cas6**; crystal structure of repeat wrapped around Cas6 [326]
Superclass C					
c	*Escherichia coli* K12 substr. W3110	F4	M2	I-E	Structure predicted, but stable; 8-nt 5' tag; cleavage by **Cas6e**, biochemical experiments [32]
d	*Thermus thermophilus* HB8	F4	M2	I-E	Structured; 8-nt 5' tag; cleavage by **Cas6e**; crystal structure of repeat hairpin in Cas6e (Cse3) [103, 274]
e	*Pseudomonas aeruginosa* UCBPP-PA14	F5	M4	I-F	Cleavage by Cas6f (Csy4); 8-nt 5' tag; crystal structure and mutational analyses of repeat hairpin in **Cas6f** [130, 131, 298]
Superclass D					
f	*Bacillus halodurans* C-125	F3	M3	I-C	Cleavage by **Cas5d**; 11-nt-5'-tag mutational analysis of hairpin structure [227]
g	*Thermus thermophilus* HB27	F37	M9	I-C	Cleavage by **Cas5d**; 11-nt-5'-tag biochemical experiments [99]
h	*Nanoarchaeum equitans* Kin4-M	-	-	I-A	Biochemical evidence to show **Cas6b** activity; 8-nt 5' tag [248]
Superclass E					
i	*Synechocystis* sp. PCC6803	-	M5	I-D & III-variant	Cleavage by **Cas6**; 8-nt 5' tag; biochemical experiments, extended structure prediction of hairpin motif [P10]
j	*Methanosarcina marzei* Gö1	F26	M13	I-B & III-B	Cleavage by **Cas6b**; 8-nt 5' tag; structure probing experiment of hairpin [P7]
k	*Clostridium thermocellum* ATCC 27405	F20	-	I-B	Biochemical evidence to show **Cas6b** activity; 8-nt 5' tag [260]
l	*Staphylococcus epidermidis* RP62A	-	M28	III-A	Cleavage by **Cas6**; 8-nt 5' tag; hairpin structure as in M28 verified by mutational analysis and sequence specificity around cleavage site [129]
m	*Methanococcus maripaludis* C5	-	M29	I-B	Cleavage by **Cas6b**; 8-nt 5' tag; biochemical experiments [260]
n	*Synechocystis* sp. PCC6803	-	M14	III-variant	Biochemical analysis of **Cmr2** implicate its involvement in either cleavage, crRNA stabilisation, or array expression regulation; 13-nt 5' tag [P10]
o	*Streptococcus pyogenes* SF370 (M1 serotype)	F35	-	II-A	Cleavage with **tracrRNA**, host **RNase III** and **Cas9**, biochemical experiments; 22-nt 5' tag [60]

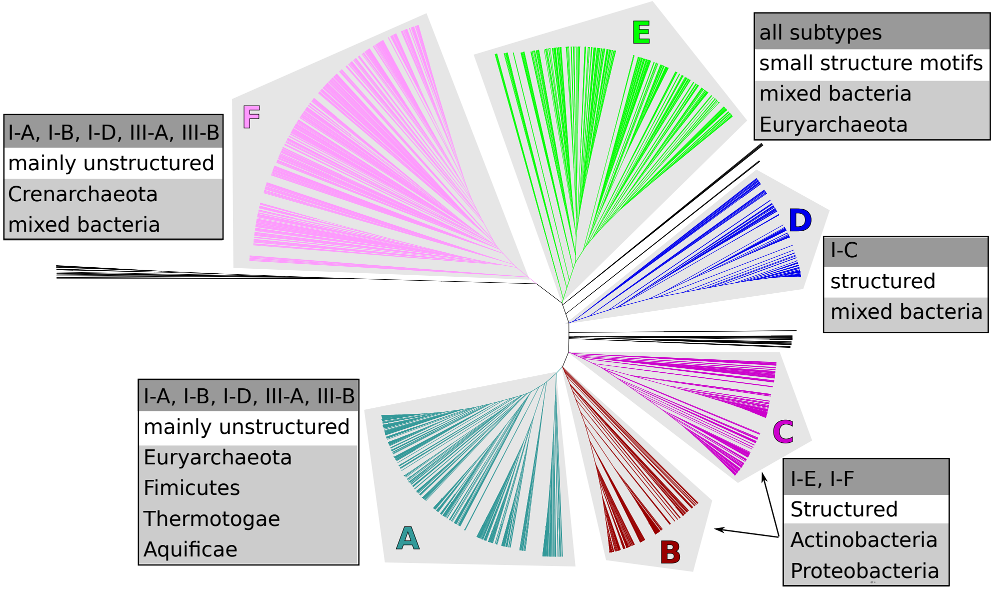


Figure 3.2. CRISPRs cluster into six major superclasses according to sequence and structure similarity. We summarised general results of our structure motif detection (i.e., structured or unstructured), Cas-subtype annotations [202], and taxonomic phyla beside each superclass. Figure taken from [P3].

3.4.1 Structure motifs fit to known cleavage sites

Most sequence families and structure motifs are associated with either bacterial or archaeal CRISPRs: only four motifs (M11, M20, M29, and M31) and one family (F20) are considerably mixed with respect to the domain (archaea or bacteria). Bacterial CRISPRs are more structured in general than those from archaea. Although structured motifs were identified for both domains, the longer, more thermodynamically stable hairpins—associated with Cas subtypes I-C, I-E, and I-F—belonged almost exclusively to bacterial CRISPRs in superclasses B–D (Figure D.11.A–C and Tables D.6–D.11). To add to the stability of such short hairpin motifs, 65 % of base pairs are Gs paired to Cs. In a closer inspection, we observed that 94 % of GC base pairs were orientated with the G towards the 3' end (Tables D.2–D.19). Such consecutive $C \to G$ base pairs form a 3' G side to the stem, which might be important for crRNA processing due to sequence specificity in this region [129, 227, 298].

In the literature, cleavage by known Cas6-like endoribonucleases (during crRNA maturation) occurs either at the 3' side of the bottom of the hairpin motif, or within the double-stranded region of the hairpin stem, usually below such a $C \to G$ base pair [32, 99, 103, 129–131, 227, 248, 260, 274, 326] [P7, P10]. The product of this cleavage is an 8-nt-long repeat tag at the 5' end of the mature crRNA (5' tag), which corresponds to the last eight nucleotides from the 3' end of the repeat sequence. Some exceptions to the 8-nt length exist [P10] [60, 99, 227, 290]. We located potential cleavage sites on our structure motifs according to published observations [32, 129, 227, 274, 298] [P7, P10]. Of all 33 structure motifs, 11 contain a potential cleavage site between two base pairs in the conserved stem of the motif of which 7 are below a $C \to G$ base pair. Another 13 motifs have a potential cleavage site at the 3' side of the bottom of the conserved stem. In Figure 3.2, we see that both of the

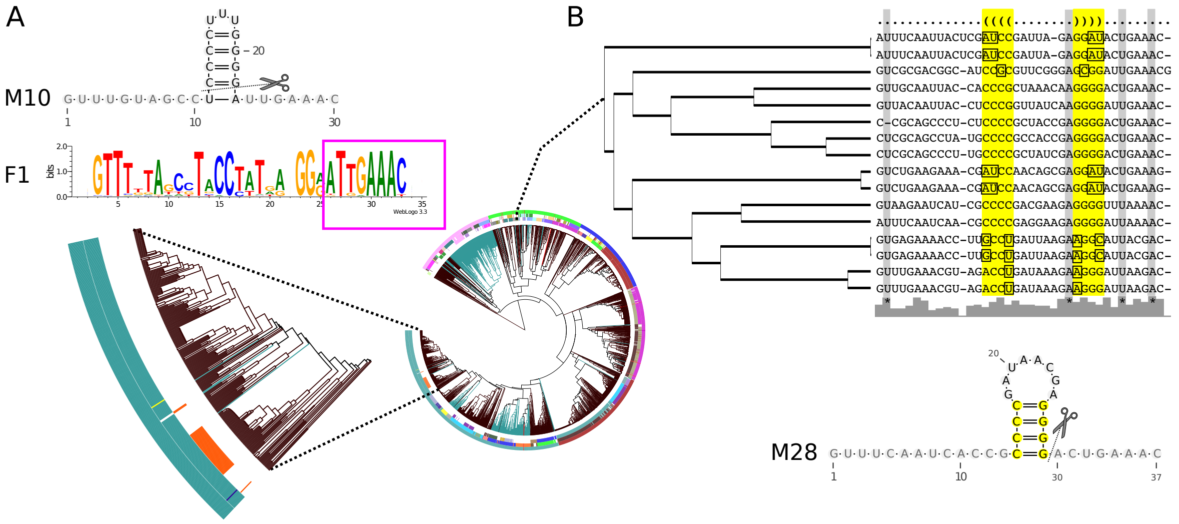

A
M10
G·U·U·U·G·U·A·G·C·C·U—A·U·U·G·A·A·A·C
1
10
20
30
F1
bits
WebLogo 3.3
B
...............((((.........))))........
AUUUCAAUUACUCGAUCCGAUUA-GAGGAUACUGAAAC-
AUUUCAAUUACUCGAUCCGAUUA-GAGGAUACUGAAAC-
GUCGCGACGGC-AUCCGCGUUCGGGAGCGGAUUGAAACG
GUUGCAAUUAC-CACCCGCUAAACAAGGGGACUGAAAC-
GUUACAAUUAC-CUCCCGGUUAUCAAGGGGAUUGAAAC-
C-CGCAGCCCU-CUCCCCGCUACCGAGGGGACUGAAAC-
CUCGCAGCCUA-UGCCCCGCCACCGAGGGGACUGAAAC-
CUCGCAGCCCU-UGCCCCGCUAUCGAGGGGACUGAAAC-
GUCUGAAGAAA-CGAUCCAACAGCGAGGAUACUGAAAG-
GUCUGAAGAAA-CGAUCCAACAGCGAGGAUACUGAAAG-
GUAAGAAUCAU-CGCCCCGACGAAGAGGGGUUUAAAAC-
AUUUCAAUCAA-CGCCCCGAGGAAGAGGGGAUUAAAAC-
GUGAGAAAACC-UUGCCUGAUUAAGAAGGCAUUACGAC-
GUGAGAAAACC-UUGCCUGAUUAAGAAGGCAUUACGAC-
GUUUGAAACGU-AGACCUGAUAAAGAAGGGAUUAAGAC-
GUUUGAAACGU-AGACCUGAUAAAGAAGGGAUUAAGAC-
M28
G·U·U·U·C·A·A·U·C·A·C·C·G·C=G·A·C·U·G·A·A·A·C
20
1
10
30
37

Figure 3.3. Highlighting the advantage of independent clustering approaches. (A) CRISPRs in the largest sequence family, F1, are mostly unstructured; however, for 50 CRISPRs also a conserved structure motif, M10, was identified. This indicates that subsets of conserved families can be structured. F1 contains the conserved 5' tag, marked with the magenta box. (B) Structure motif M28 shows no sequence conservation, but a conserved structure (base pairs are highlighted in yellow). The many compensatory base pairs are marked in the alignment with squares. This structure has been verified via mutational analyses in [129]. Potential cleavage sites are indicated as observed in the literature [32, 99, 103, 129–131, 227, 248, 260, 274, 326] [P7, P10]. Figure taken from [P3].

Cas subtypes I-E and I-F are split across the two superclasses B and C. The splitting of these subtypes is due to a single repeat-structure feature: The hairpin motifs are closer to the 3' end of the CRISPRs in superclass B, resulting in a cleavage site within the stem. In superclass C, the cleavage site is at the bottom of the hairpin motif. In accordance with previously mentioned literature, the cleavage sites are below a $C \rightarrow G$ base pair in both superclasses. Aside from this difference in position, the hairpin structures associated with either Cas subtypes I-E or I-F are very similar. See Figure D.11 for details.

3.4.2 Patterns of conservation in sequence families

When inspecting the family sequence logos, we see different patterns of *sequence* conservation (Figure D.11 and Tables D.2–D.19). We highlight these differences using four selected examples: First, CRISPRs associated with the Cas I-E subtype show a high conservation of Gs and Cs that form the base pairs of the hairpin motif. Second, CRISPRs associated with the I-F subtype are well-conserved across the entire repeat sequence and contain fewer consecutive Cs and Gs (Figure D.11.A–B). Third, CRISPRs associated with the Cas I-C subtype show a higher conservation at the bottom of the hairpin stem and in the single-stranded 5' and 3' ends, which suggests that the top of the stem and the hairpin loop is likely insignificant for the binding affinity (Figure D.11.C); this conservation pattern is well-supported by mutation experiments in the type I-C system in *B. halodurans* C-125 where crRNAs were still processed with a truncated upper stem and mutated hairpin loop, but processing was sequestered by mutations at the bottom of the stem or by the removal of the unpaired 3' end [227]. Fourth, in Figure 3.3, we marked the well-conserved 8-nt-long 5' tag, $AUUGAAA(C/G)$. Out of our 40 sequence families, 17 (∼40 %) show a conservation of exactly this sequence tag; others contain minor deviations. Interestingly, bacterial superclasses B and C do not show this tag, whereas it is highly conserved throughout the other bacterial superclass D and in almost all archaeal families (9 out of 12). We hypothesise that these patterns of conservation give a good indication of differences in binding affinities for specific Cas proteins in various CRISPR-Cas systems.

3.4.3 Sequence families and structure motifs provide independent information about evolution

Structured ncRNA families cannot be identified by sequence conservation alone, since standard alignment tools fail when the pairwise sequence identity is below 60 % [96]. We see the same tendency for structured and unstructured repeats in our data: The `CRISPRmap` tree shows different patterns of overlap between sequence families and structure motifs that we identified by independent clustering approaches (Figure 3.1). In Figure 3.3, we highlight two overlap patterns. First, in superclass A, the largest family, namely F1, is mainly unstructured. For a subset of these CRISPRs, however, we identified a thermodynamically stable hairpin motif (M10) with four, consecutive $C \rightarrow G$ base pairs; these CRISPRs are clearly structured. Second, in superclass D, we found a conserved hairpin motif (M28), also with four, consecutive $C \rightarrow G$ base pairs and a large 8-nt hairpin loop that was verified by mutational analyses in a type III-A system in *Staphylococcus epidermidis* RP62A [129]; this motif does not show enough sequence conservation to be detected as a sequence family. Both M10 and M28 would not have been identified with the approach used in [180], in which consensus structures were calculated from (entire) sequence families. In addition, we observe cases where a structure motif corresponds almost fully to a sequence family, e.g., M1 with F2 and M2 with F4. Nevertheless, individual members of the sequence families were not predicted to form the associated consensus structure: this may indicate a degenerate and non-functional CRISPR-Cas system, or one that has evolved to function with a different or no repeat structure.

3.4.4 A subset of Cas subtypes are weakly linked to repeat and Cas1 evolution

From the literature, we know that Cas1 is strongly linked to repeat evolution [98,144]. This link could be verified for our large-scale dataset (Figure 3.4.A). To do this, we clustered associated Cas1-protein sequences and the results fit well with all superclasses, except superclass E[1] (Figure 3.4).

[1] There are two observations which indicate that superclass E contains only partial data: conserved sequence families and structure motifs are smaller and most CRISPRs show little to no conservation, and in Section D.10, we identified that a large number of CRISPRs from metagenomic data were assigned to this superclass that potentially form conserved classes.

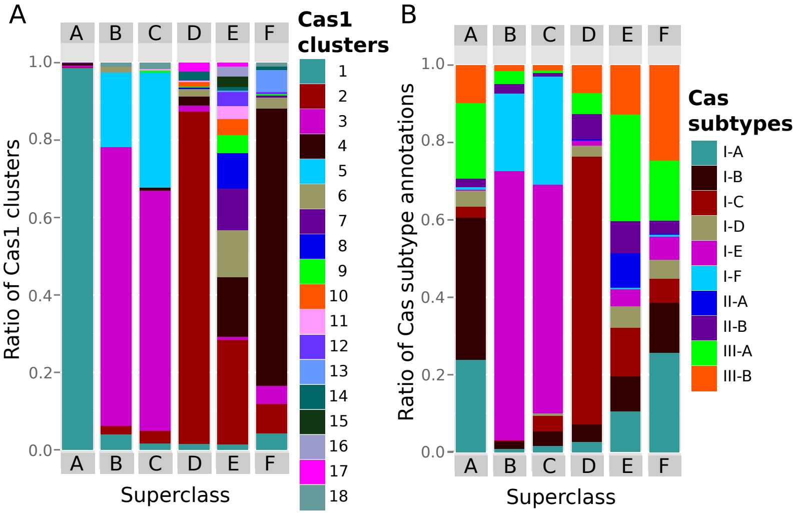


Figure 3.4. Relative ratios of Cas1 sequence clusters and Cas-subtype annotations per superclass. (A) Cas1 sequence clusters correspond well to the superclass and thus the CRISPRmap tree with the exception of superclass E; superclass E is very diverse in both repeat and associated Cas1 conservation and it probably contains only partial data. (B) Bacterial CRISPRs that are assigned to well-defined structure motifs are associated with subtypes I-C, I-E, and I-F in superclasses B–D and are strongly linked to both repeat and Cas1-sequence similarities (i.e., CRISPR evolution). Superclass A and F contain both bacterial and archaeal CRISPRs (many are unstructured), which are loosely associated with the remaining type I and both type III subtypes. These subtypes do not correspond to Cas1 and repeat evolution and are likely comprised of interchangeable protein complexes or modules. The diversity of superclass E is also reflected by the mixture of all subtypes. In addition, the majority of type II CRISPRs are also located in this region. Figure taken from [P3].

From our data, we observed that the linkage between Cas subtype and repeat evolution is not so clear: subtypes I-C, I-E, and I-F correlate well with repeat (and thus Cas1) conservation, whereas, the remaining type I and both type II Cas subtypes are only weakly linked (Figure 3.4). The bacterial superclasses B, C and D contain well-defined structure motifs and sequence families (Figure 3.1 and Tables D.2–D.19), which are associated with subtypes I-E and I-F (superclasses B and C) and I-C (half of superclass D). Superclasses A and F contain both bacterial and archaeal CRISPRs—most of which are unstructured—and although they also fit well to the Cas1 clusters, the annotated Cas subtypes are a diverse mixture of the remaining type I subtypes (I-A, I-B, and I-D) and both type III subtypes (Figure 3.4). In accordance with the diversity seen in the Cas1 analysis, superclass E also contains all subtypes.

The following may explain why there is a co-occurrence of type I and type III subtypes. First, these subtypes are composed of interchangeable modules as was previously suggested for archaeal systems [98, 286]. In such cases one would expect Cas proteins from different subtypes to be able to process similar repeat sequences. Two examples in the literature support this theory: (1) a Cas6 (Cas6b) protein that can process both type I-B systems in *Methanococcus maripaludis* C5 and *Clostridium thermocellum* ATCC 27405 [260]; and (2) two CRISPRs in *Methanosarcina marzei* Gö1 with near-identical repeats are associated with different subtypes I-B and III-B [P7]. Also, many sequence families and structure motifs co-occur with multiple, or a mixture of, subtypes (see Tables D.2–D.19). The co-occurrence of subtypes is wide-spread in archaea and bacteria. In general, an exchange of protein modules would require compatible repeat sequences and structures. The only similarity observed in CRISPRs associated with mixed subtypes is the conserved 5' tag—*AUUGAAAC/G*, or a slight variation thereof. In comparison, repeats associated with the bacterial subtypes I-E and I-F do not contain this tag. Second, additional or unknown Cas proteins are required to achieve a sub-classification of Cas subtypes that is more compatible with repeat conservation. Most likely, the truth lies in a combination of both explanations. Finally, we observed that subtypes I-A, I-B, I-D, III-A, and III-B are more enriched in extremophiles, e.g., thermophiles (Figure D.6). Perhaps organisms living in extreme conditions benefit from a mechanism that involves a rapid transfer and a reconfiguration of CRISPR-Cas systems.

3.4.5 CRISPRs in Euryarchaeota are closer to bacterial systems than ones in Crenarchaeota

97 % of the archaeal CRISPRs originate from two phyla: 380 from Euryarchaeota and 245 from Crenarchaeota. In the `CRISPRmap` tree (Figure 3.1 and Figure D.4), we observe a clear separation of these two CRISPR groups. 60 % of CRISPRs from Euryarchaeota and 96 % from Crenarchaeota cluster into superclasses A and F, respectively. In superclass A, the euryarchaeal and bacterial CRISPRs are associated with Cas1 proteins that cluster into the same Cas1-cluster-1, i.e., these Cas1 sequences are evolutionarily close (Figure 3.4). In contrast, CRISPRs from Crenarchaeota are located almost exclusively in a sub-region of superclass F and are associated with the separate Cas1-cluster-4 (Figure D.4).

3.4.6 Evidence of horizontal transfer

As previously mentioned, the majority of archaeal and bacterial CRISPRs are distinctly separated in the `CRISPRmap` tree (Figure 3.1). This is consistent with a rare exchange of genetic material between archaeal and bacterial systems [98, 287]. Nevertheless, we observed a few instances where archaeal repeats are located in a bacterial-dominated region of the tree and vice versa (see Appendix D.1.1 for more details). With one exception, it is assumed that all cases involved a transfer of the CRISPR-Cas system from bacteria to archaea; archaea have also been shown to uptake bacterial and eukaryotic DNA as spacers [31]. Figure D.9 gives examples of archaea that contain full bacterial CRISPR-Cas systems where a strong conservation of the structure motif is supported by multiple compensatory base pair mutations. In addition, the archaeal CRISPRs are associated with the complete set of proteins from the bacterial subtypes I-C and I-E.

The transfer of genetic material between prokaryotes often occurs via plasmids, however, in Figure D.9 all horizontally transferred systems in the archaea are located on chromosomes and not on plasmids. In fact, overall only 7 % of over 1,300 plasmids analysed contained a CRISPR array. Therefore, it is unlikely that the dominant mechanism of transferring CRISPR-Cas systems between organisms is via plasmids.

3.5 The CRISPRmap web server

The `CRISPRmap` web server enables easy access to our data and allows scientists to compare the conservation of individual repeats. Repeats are entered in FASTA format and the web server automatically assigns them to our classification system; previously unknown repeats are assigned to existing families and/or motifs, if possible. Non-conserved input sequences remain unassigned, but are still located according to their relative similarity in the tree. Furthermore, if the correct orientation of the input repeats is unknown, the user can request to predict the orientations to ensure that they are consistent with our data.

The user of our `CRISPRmap` web server can enter up to 300 CRISPR sequences in FASTA format and indicate whether the correct orientation is unknown and requires prediction. We use a multi-step procedure that has been optimised for speed to assign the given repeats to our structure motifs and sequence families. Further details are given in Appendix D.1.1.

All data and the web server are available under `http://rna.informatik.uni-freiburg.de/CRISPRmap`.

3.5.1 Comparison of published CRISPR systems

We employed the `CRISPRmap` web server to verify our methods by comparing results with CRISPR-Cas systems where the crRNA maturation mechanism has been characterised by wet-lab experiments. Published information was consistent with our identified structure motifs, subtype annotations, and our predicted orientations (Table 3.2). The previously mentioned co-occurrence of subtypes I-A, I-B, I-D, and type III is verified in part by the published systems in superclass E (see Table 3.2, IDs i-n). Further comparisons are given in Section 4.1.

3.6 Conclusion

We provide a comprehensive analysis of CRISPR structure and sequence conservation based on the largest dataset of repeat sequences available. We show extensively that our methods are well-suited to identifying many characteristics of CRISPR-Cas systems: e.g., cleavage sites, patterns of RNA structure motifs and sequence conservation, the link between evolution of CRISPRs and associated Cas subtypes, and the horizontal transfer of such systems. On the one hand, specific conservation patterns can be combined with published data to make assumptions about CRISPRs belonging to the same sequence families or structure motifs. On the other hand, the CRISPRmap overview can be used to find potentially novel CRISPR-Cas systems that are highly divergent from the rest. User-based queries on our data enable more informed choices on future hypotheses in CRISPR-Cas research.

CHAPTER 4

Applications and limitations of CRISPRmap

This chapter highlights possible applications of the CRISPRmap web server and data from Chapter 3. Both applications and limitations of previous work are discussed using published examples.

4.1 Application of CRISPRmap to single systems

About 30 % of bacteria and 70 % of archaea contain at least one CRISPR-Cas system. When capturing similarities and exploring the diversity of these systems, it is impossible to characterise every single system. Therefore, representatives of conserved groups are selected for further analysis. Conserved groups can either be determined by associated Cas proteins, given by published Cas-subtype annotations [123, 201, 202, 324] or by CRISPR conservation. In the previous chapter, we determined that Cas subtypes are only weakly linked to CRISPR (repeat) conservation (Section 3.4.4). Therefore, additional representatives for further characterisation may also be chosen according to CRISPR sequence and structure conservation.

In collaboration with the lab of Prof. Dr. Anita Marchfelder, we studied the CRISPR-Cas systems encoded in *Haloferax volcanii* H119. We were particularly interested in the processing mechanism, which involves CRISPR expression and crRNA maturation (Section 2.2.2). During crRNA maturation, CRISPR RNA is generally processed by a Cas endoribonuclease of the Cas6 family [32, 36, 103, 126, 129, 130, 155, 248, 260, 274, 326] and [P7, P10]. Assuming that both Cas6 and CRISPR coevolve, one can simply apply the CRISPRmap web server to compare the *H. volcanii* Cas6 with those previously published. Using this approach, we observed that CRISPRs encoded in Haloferax species are on a distinct branch of the CRISPRmap tree that is clearly distant from all previously studied systems (Figure D.12).

Using this as evidence that the Cas6 proteins in *H. volcanii* may give additional insights into Cas6-based mechanisms, members of Prof. Dr. A. Marchfelder's lab performed experiments to characterise Cas6 function in *H. volcanii* [P2].

4.2 Novel CRISPRs in metagenomic data indicate a vast spectrum of diversity

A valuable source of new CRISPR-Cas systems are metagenomic studies of multiple, often novel, prokaryotes. Recently, 150 CRISPR arrays were identified in the bacterial metagenome from different sites in the human body [256]. We applied `CRISPRmap` to quickly determine the conservation of these CRISPRs: only 38 % and 29 % were assigned to our structure motifs or sequence families, respectively. Notably, 50 % of the metagenomic CRISPRs were assigned to the diverse superclass E where most remained unassigned to either a structure motif or sequence family. However, in Figure D.10, many of these repeats cluster together to potentially form new classes of motifs and families. Two CRISPRs fall into the euryarchaeal region in superclass A, despite the fact that archaea are rarely associated with human microbiomes [256].

A similar study was performed in [107] where the authors used `CRISPRmap` to classify 233 CRISPRs identified in the human gut metagenome. Similar to the previous study, these CRISPRs belonged predominantly to the superclasses with little sequence conservation. These results highlight the fact that even with the large-scale analysis performed in this work, we still do not know the full extent of CRISPR-Cas diversity. Therefore, the dynamic nature of our web server—in the fact that it allows the classification of newly sequenced CRISPRs to be assigned to existing sequence families and structure motifs—is particularly useful.

4.3 Limitations of the `CRISPRmap` web server

The `CRISPRmap` version 0.1, as published in [P3], is limited mainly by the following two factors: despite the attempts at orientation prediction, many CRISPRs are still in the incorrect orientation [22] and structure motifs were limited to constitute at least three base pairs. The first factor leads to incorrect clustering and the second to missing data in the tree. Therefore, individual analyses of conservation can still be beneficial.

Although `CRISPRmap` could be automatically updated to include newly sequenced CRISPR data, this data could considerably change results. Therefore, to limit confusion, the `CRISPRmap` web server requires regular updates to capture the full diversity of sequenced CRISPR-Cas systems, and also must adapt to potential changes in Cas protein and subtype annotations.

4.3.1 CRISPRmap cannot detect hairpin motifs with only three base pairs

In [P5], we characterised the CRISPR-Cas systems in the thermophilic archaeon *Haloferax volcanii* H119. According to CRISPRmap and previous work from Kunin and colleagues [180], the three CRISPRs encoded in *H. volcanii* are unstructured as no conserved structure was identified. We analysed the folding potential of each single repeat in all three of the CRISPR RNAs as a function of the surrounding spacer sequences. According to these analyses, all *H. volcanii* CRISPRs called C (located on the chromosome), P1 and P2 (both located on a plasmid) share a minimal three base-pair stem loop [P5]. A comparative approach, using CRISPR repeat sequences from other haloarchaeal genomes, corrobated the significance of the minimal hairpin motif as it was conserved in all analysed haloarchaea (Figure 4.1). This conserved structural motif is generally surrounded by additional base pairs within the repeat and contains three consecutive $C \rightarrow G$ base pairs for stability.

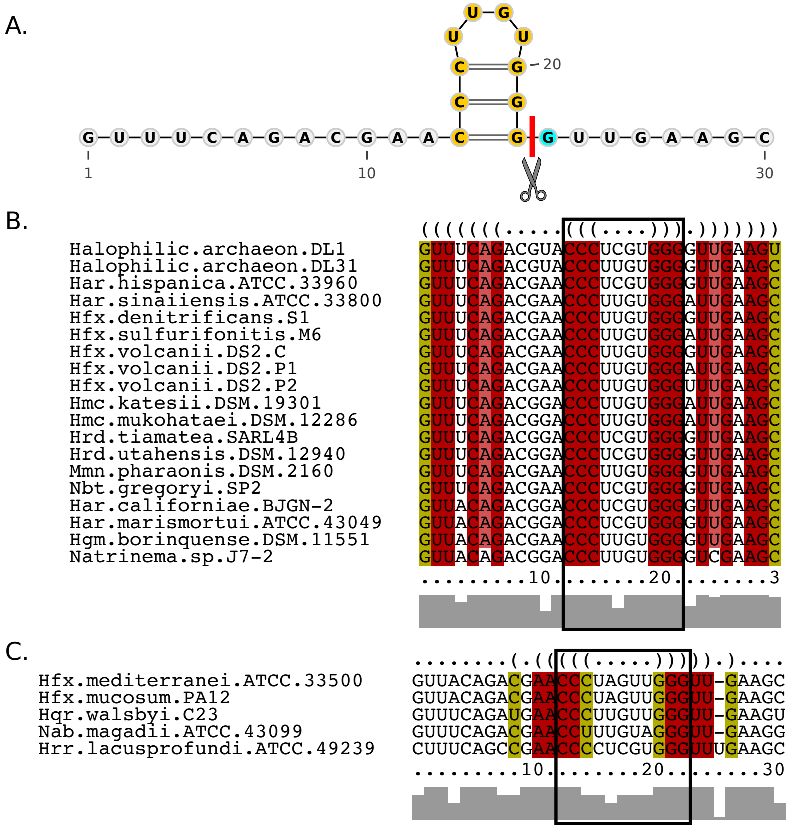

A.
G U U U C A G A C G A A C C C U U G G G U U G A A G C
U G
1
10
20
30
B.
((((((((.....(((....))).)))))))
Halophilic.archaeon.DL1 GUUUCAGACGUACCCUCGUGGGGUUGAAGU 30
Halophilic.archaeon.DL31 GUUUCAGACGUACCCUUGUGGGGUUGAAGC 30
Har.hispanica.ATCC.33960 GUUUCAGACGAACCCUCGUGGGGUUGAAGC 30
Har.sinaiiensis.ATCC.33800 GUUUCAGACGAACCCUUGUGGGAUUGAAGC 30
Hfx.denitrificans.S1 GUUUCAGACGAACCCUUGUGGGGUUGAAGC 30
Hfx.sulfurifonitis.M6 GUUUCAGACGAACCCUUGUGGGAUUGAAGC 30
Hfx.volcanii.DS2.C GUUUCAGACGAACCCUUGUGGGGUUGAAGC 30
Hfx.volcanii.DS2.P1 GUUUCAGACGAACCCUUGUGGGAUUGAAGC 30
Hfx.volcanii.DS2.P2 GUUUCAGACGAACCCUUGUGGGUUUGAAGC 30
Hmc.katesii.DSM.19301 GUUUCAGACGGACCCUUGUGGGAUUGAAGC 30
Hmc.mukohataei.DSM.12286 GUUUCAGACGGACCCUUGUGGGAUUGAAGC 30
Hrd.tiamatea.SARL4B GUUUCAGACGGACCCUCGUGGGGUUGAAGC 30
Hrd.utahensis.DSM.12940 GUUUCAGACGGACCCUCGUGGGGUUGAAGC 30
Mmn.pharaonis.DSM.2160 GUUUCAGACGAACCCUUGUGGGGUUGAAGC 30
Nbt.gregoryi.SP2 GUUUCAGACGAACCCUCGUGGGGUUGAAGC 30
Har.californiae.BJGN-2 GUUACAGACGGACCCUCGUGGGGUUGAAGC 30
Har.marismortui.ATCC.43049 GUUACAGACGGACCCUCGUGGGGUUGAAGC 30
Hgm.borinquense.DSM.11551 GUUACAGACGAACCCUUGUGGGGUUGAAGC 30
Natrinema.sp.J7-2 GUUACAGACGGACCCUUGUGGGGUCGAAGC 30
.........10.........20........3
C.
........(.((((.....))))).)....
Hfx.mediterranei.ATCC.33500 GUUACAGACGAACCCUAGUUGGGUU-GAAGC 30
Hfx.mucosum.PA12 GUUACAGACGAACCCUAGUUGGGUU-GAAGC 30
Hqr.walsbyi.C23 GUUUCAGAUGAACCCUUGUUGGGUU-GAAGU 30
Nab.magadii.ATCC.43099 GUUUCAGACGAACCUUUGUAGGGUU-GAAGG 30
Hrr.lacusprofundi.ATCC.49239 CUUUCAGCCGAACCCCUCGUGGGUUUGAAGC 31
.........10.........20.........30

Figure 4.1. A small minimal hairpin structure motif is conserved across 22 haloarchaeal species. (A) Part of the predicted structure for the repeat from locus C is conserved throughout the haloarchaeal species (highlighted in yellow). The red line corresponds to the determined cleavage site just upstream of the 5' crRNA tag. The G nucleotide that is cyan in colour corresponds to the 23rd nucleotide, which is an A at locus P1 and a U at locus P2. (B and C) Multiple sequence alignments generated by LocARNA; the red columns correspond to conserved base pairs and the mustard yellow columns correspond to the presence of a compensatory base pair that conserves the consensus structure. The conserved structural motif from (A) is surrounded by the black box. (B) The larger group of haloarchaea with the conserved motif and a 4-nucleotide hairpin loop. (C) The smaller group with a 5-nucleotide hairpin loop. The conserved CG stem-loop motif is surrounded by stabilising base pairs in both groups. Figure taken from [P5].

The three-base-pair hairpin motif could not be confirmed *in vitro* experiments, however due to growth in high salt conditions, it may still be present *in vivo* [P5]. Furthermore, Nam and colleagues determined that a minimum hairpin of three base pairs is sufficient for recognition and cleavage by the Cas5d protein in *Bacillus halodurans* [227]. Changing the minimum number of base pairs required for a structure motif in `CRISPRmap` from four to three did not generate acceptable results. We observed a large influx in the number of structure motifs identified with low sequence identities. Therefore, we assumed that motifs with only three base pairs frequently occur by chance. In the case of the Haloferax CRISPRs, however, sequence conservation is high (see Figure 4.1) and this motif might be required for recognition by Cas6.

4.3.2 Improved prediction of CRISPR orientation

CRISPRs are transcribed and processed into mature crRNAs generally from only one strand. The transcribed strand determines the orientation of the CRISPR sequence, which is important to know for evolutionary analyses. From our previous work and from the literature, we observed two factors that could be indicative of the correct CRISPR orientation. First, the 5' and 3' ends of CRISPRs were generally more conserved than the middle section. For example, the 8-nt tag *AUUGAAAG/C* that remains at the 5' end of the mature crRNA was conserved in 40 % of all `CRISPRmap` sequence families (Section 3.4.2). Second, the CRISPR locus contains more mutations towards the 3' end of the repeat-spacer array, since these are usually the oldest; adaptation generally occurs at the 5' end, adjacent to the leader [8]. Third, archaeal CRISPRs are rich in *A*s in the correct orientation and are considerably depleted in poly(T) signals (more than three *T*s in a row) in our `REPEATS` from Chapter 3; poly(T) regions are signals of transcription termination in archaea [P9]. Therefore, we extended the initial graph model for predicting orientations from `CRISPRmap` version 1.0 to include mutational and positional information as well as just the sequence. Prediction accuracies increased by over 10 % AUROC. Correctly predicted CRISPR orientations lead to a better clustering using the `CRISPRmap` pipeline and the web server was thus updated to version 2.0 [P1]. The updated `CRISPRmap` web server (version 2.0) is available at `http://rna.informatik.uni-freiburg.de/CRISPRmap`— and so is the standalone orientation software (`CRISPRstrand`). Details of the orientation-prediction method are described in [P1][1].

4.4 Conclusion

`CRISPRmap` provides a comprehensive overview of CRISPRs from published systems and it has already been successfully applied here [P2] and in other published work [256]. Despite some limitations, the improved prediction of CRISPR orientation has enhanced the quality of the `CRISPRmap` data and it is currently the only application available that performs an automated classification of CRISPR conservation. Regular updates of the `CRISPRmap` web server are planned.

[1] The majority the orientation-prediction work was performed by Omer S. Alkhnbashi and was therefore kept to a brief summary here. I contributed to the development of the underlying methods.

Part III

Analysis of non-coding RNA expression

Part III: Analysis of non-coding RNA expression

Self-expression must pass into communication for its fulfilment—Pearl S. Buck

Before RNA can perform its function, it needs to be expressed from DNA. Two factors are essential to the precise regulatory function of an ncRNA: (1) its mature form and (2) its abundancy in the cell. Transcriptome data, e.g., as derived from `RNA-seq` experiments, can provide detailed insight into ncRNA abundances and their processing.

In Chapter 5, we performed in-depth analyses of `RNA-seq` data to determine crRNA expression in organisms that encode CRISPR-Cas systems and to assess how they were processed from the transcribed CRISPR RNA. After expression, crRNAs are stabilised to prevent their immediate degradation. Therefore, in this part, we investigated attributes that might influence crRNA stability. The majority of this chapter was presented in [P10].

CHAPTER 5

Expression and processing of mature crRNAs

Active CRISPR-Cas–based defence against invaders requires a stable population of crRNAs to target and degrade foreign genetic material (Section 2.2.2 gives an introduction to CRISPR-Cas immune systems). A CRISPR array is first transcribed and then processed into single-spacer units, crRNAs, by endoribonucleases (usually by members of the CRISPR-associated Cas6 protein family) that cleave the RNA at each repeat instance in the array [2, 307, 337]. Among other factors, expression of mature crRNAs depend on their efficient and accurate processing and their subsequent stabilities in the cell. In *Haloferax volcanii*, we observed that not all crRNAs, derived from one expressed CRISPR array, lead to a successful defence reaction [P5]. The assumption is that unsuccessful crRNAs are either not correctly processed or not stably integrated into their effector complex. Once in the effector complex, crRNAs are protected from degradation and enable the destruction of foreign material via base pairing to their complementary protospacers [P2]. Hence, active CRISPR-Cas systems in any organism are first characterised by establishing accurate crRNA expression.

In the following analyses, we inspected and processed `RNA-seq` data to investigate crRNA maturation and stability in a model cyanobacterium *Synechocystis* sp. PCC6803. Results were published in [P10]. Analagous investigations were performed for further organisms and published in [P5, P8].

The chapter starts with an overview of the CRISPR systems encoded in *Synechocystis.* A brief summary of associated Cas proteins gives insight into putative endoribonucleases that are involved in the crRNA maturation process. The aim is to complement the wet-lab investigation of associated Cas proteins with a computational, in-depth analysis of `RNA-seq` data. We capture processing intermediate and mature crRNA transcripts and accentuate similarities and differences of each CRISPR locus. To this end, after establishing the CRISPR-Cas systems, the methods applied to the `RNA-seq` data analysis and the results obtained from these data are presented. All wet-lab experiments were performed by members of Prof. Dr. W. R. Hess's group.

5.1 CRISPR-Cas systems encoded in *Synechocystis*

The plasmid pSYSA of *Synechocystis* sp. 6803 is a large, extrachromosomal element that is almost entirely devoted to three different CRISPR-Cas systems, CRISPR1–3, located on the forward strand; no systems are encoded in the chromosome, which is rare according to observations from our `CRISPRmap` data (see Section 3.1). Each repeat-spacer array is adjacent to a distinct set of associated cas genes (see Figure D.13). Among CRISPR1 genes are homologs to *cas3* (*slr7010*) and *csc3/cas10d* (*slr7011*), which serve as markers of CRISPR subtype I-D [202]. In contrast, CRISPR2–3 resemble type III systems, indicated by the presence of *cmr2/cas10* homologs. Other subtype-specific markers such as *csm2* or *cmr5*, however, are missing [202].

According to the previously published plasmid sequence [158][1], CRISPR1–3 consist of 49, 56 and 38 repeat-spacer units (each with an additional final repeat), respectively. The spacer sequences differ in length from 31–47 nt, and with the exception of a few identical spacers within CRISPR1 and CRISPR2, they are all unique. Identical single repeat-spacer units and pairs of two adjacent identical repeat-spacer units appear in a consecutive manner in CRISPR1 and CRISPR2.

5.1.1 Experimental analysis to identify the processing endoribonucleases

In type I and III CRISPR-Cas systems, the large and diverse protein family, the Cas6 endoribonucleases, have been shown to cleave CRISPR arrays at repeat instances [2,307,337]. In general, a Cas6 endoribonuclease binds specifically to its associated repeat and cleaves it such that an 8-nt repeat handle remains on the mature crRNA; for many CRISPRs, the binding motif is a small hairpin [32,103,129,131,227,274,298], see also Chapter 3.

Three potential *cas6* genes are located on pSYSA: *slr7014* (*cas6-1*) at the CRISPR1 locus; and both *slr7068* (*cas6-2a*) and *sll7075* (*cas6-2b*) at the CRISPR2 locus (Figure D.13). No *cas6* homolog is associated with CRISPR3. The pairwise similarity between the encoded protein sequences and their similarity to the functionally characterised Cas6 homolog of *Pyrococcus furiosus* [326] is very low, ranging between 6–17 % identical amino acid residues. Knock-out mutations of the three *cas6* genes and subsequent Northern analyses showed that the knock-out of *cas6-1* and *cas6-2a* affect the accumulation of crRNA transcripts from CRISPR1 and CRISPR2, respectively. This is in agreement with both their locations immediately 5' of the respective CRISPRs [P10]. Whereas the Δ*cas6-1* knock-out mutant showed a complete loss of CRISPR1 RNA accumulation, the Δ*cas6-1* mutant accumulated CRISPR2 RNA to > 200 nt, but the smaller, mature crRNA transcripts were not observed [P10]; both mutants did not affect RNA accumulations at the other CRISPR loci *in vivo*. Knocking out the expression of *cas6-2b* did not affect crRNA accumulation of any of the three loci.

In [P8], it was confirmed experimentally that Cas6-1 binds and cleaves both CRISPR1 and CRISPR2 at the positions indicated *in vitro*. This implies that the Cas6-1 protein is specific

[1] The plasmid pSYSA sequence is available in the RefSeq databank with the accession number NC_005230.

to the CRISPR1 locus *in vivo*, but is capable of binding and cleaving other CRISPR loci *in vitro*. Although CRISPR3 is not associated with a Cas6 protein and is not affected by the other encoded Cas6 proteins, we observed that a knock-out mutation of the *cmr2* gene caused a complete loss of CRISPR3 RNA transcripts. Although it has been predicted that the Cmr2 protein contains a nuclease domain, we could not verify whether it functions as an endoribonuclease or merely contributes to the stability of processed crRNAs. In summary, these experimental results point to three distinct processing mechanisms for each of the three systems encoded in *Synechocystis*.

5.1.2 CRISPR structure motifs

In Part IV, Chapter 7, we developed a method to predict the most stable structure motif for a CRISPR repeat sequence by considering all repeat instances across the CRISPR array; thus we effectively incorporate the influence of the surrounding context sequence into the prediction of the repeat structure. We applied this approach to identify the potential binding motifs of the Cas6 proteins in *Synechocystis* and present the best structure results in Figure 5.1. The repeats of all three CRISPRs were able to form characteristic hairpin structures. The hairpin motifs for CRISPR1–2 are very similar with identical 8-nt repeat tags. In accordance with the *in vitro* results where it was observed that Cas6-1 was able to cleave both repeats, the similar regions in CRISPR1–2 likely represent the binding site for the endoribonuclease. The hairpin motif for CRISPR3 is distinctly different to the other two, with its larger loop size and location closer to the 5' end of the repeat.

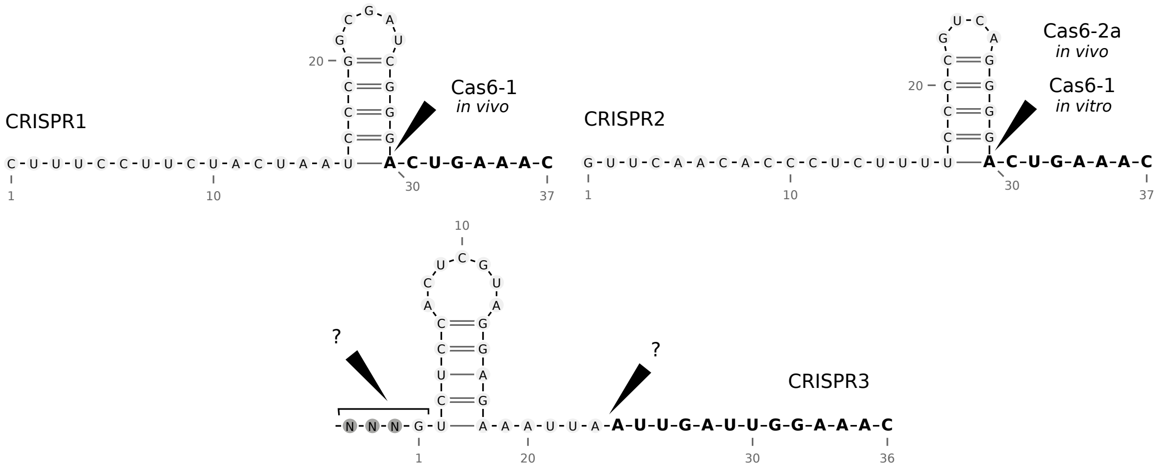

CRISPR1
Cas6-1
in vivo
C-U-U-U-C-C-U-U-C-U-A-C-U-A-A-U—A-C-U-G-A-A-A-C
1
10
20
30
37
CRISPR2
Cas6-2a
in vivo
Cas6-1
in vitro
G-U-U-C-A-A-C-A-C-C-C-U-C-U-U-U-U—A-C-U-G-A-A-A-C
1
10
20
30
37
CRISPR3
?
?
-N-N-N-G-U—A-A-A-U-U-A-A-U-U-G-A-U-U-G-G-A-A-A-C
1
10
20
30
36

Figure 5.1. CRISPR hairpin structures in *Synechocistis*. Predicted CRISPR repeat structures using our CRISPR-specific prediction approach that includes influencing context sequences (Part IV,Section 7.3). The black wedges indicate cleavage sites derived from the `RNA-seq` data and the 5' repeat sequence tag of the mature crRNAs is highlighted in bold. The 5' tags for CRISPR1 and CRISPR2 had the frequently published length of 8 nt [32, 99, 103, 129–131, 227, 248, 260, 274, 326]. CRISPR3 was cleaved twice, first at the end of the spacer and second in the middle of the repeat leaving a novel-length, 13-nt tag. Figure adapted from [P10].

5.2 RNA-seq data preparation

CRISPR expression and processing in *Synechocystis* sp. PCC6803 were analysed using two `RNA-seq` datasets (datasets A and B). In fact, the laboratory-specific substrain "PCC-M" was used in all wet-lab experiments. The cDNA libraries for both datasets were prepared by vertis Biotechnologie, Germany (http://www.vertisbiotech.com/). The exact experimental conditions for both `RNA-seq` datasets A and B are described in [P10] and [218], respectively. For generating cDNA libraries for sequencing, the RNA transcripts in dataset A were ligated with a poly(A) tail. To understand subsequent terminology, a "read" is a single sequence, which was produced during the sequencing process. Subsequent to sequencing, dataset A was nearly 200 times larger than dataset B with 33,357,164 reads in contrast to only 169,360 reads in dataset B. In addition, most reads in dataset A were of length 100 nt, whereas dataset B contained many short reads with many only 18 nt (data not shown), which look like an accumulation of degradation products. Note that no size selection of the purified RNA transcripts was performed prior to sequencing.

5.2.1 Mapping the RNA-seq data

The mapping of dataset B was performed by Mitschke and colleagues and is described in [218]. Here, we concentrate on mapping dataset A, which was sequenced by an Illumina HiSeq 2000 machine. Using the `FASTQC` analysis tool, we observed an increasingly poor sequencing quality towards read ends in this dataset, possibly due to the poly(A) tails and subsequent adapter sequences (see Figure D.14). Therefore, in a pre-processing step, the reads were trimmed with respect to their sequencing quality using the `fastq_quality_trimmer` program from the `FASTX-Toolkit` version 0.0.13 with the options `-t 13 -Q 33`. The `-Q` option is necessary, because the quality scores are used with an ASCII offset of 33 according to the Sanger format. In this way, nucleotides were trimmed if they had a quality below 13, which roughly corresponds to an estimated probability of p≥0.5 that a base call is incorrect [50]. Subsequent to trimming, the dataset was mapped with Segemehl [141] version 0.1.3 with the options `-polyA -prime3 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT'` for clipping the poly(A) tail and the 3' Illumina sequencing adapter. Following this procedure, we could successfully map approximately 98 % of the original reads to the genome. The over 30 million reads indicated good coverage of the transcriptome.

To explore the `RNA-seq` results and visualise read profiles, we used the Integrative Genomics Viewer (IGV) version 2.0.3 [265].

5.3 CRISPR RNA is highly abundant

In previous work, Mitschke and colleagues mapped all transcription start sites (TSS) in the *Synechocystis* genome [218]. According to their data, the precursor RNAs for CRISPR1–3 originate from one TSS each. The locations of the TSS result in transcribed 5' leaders of

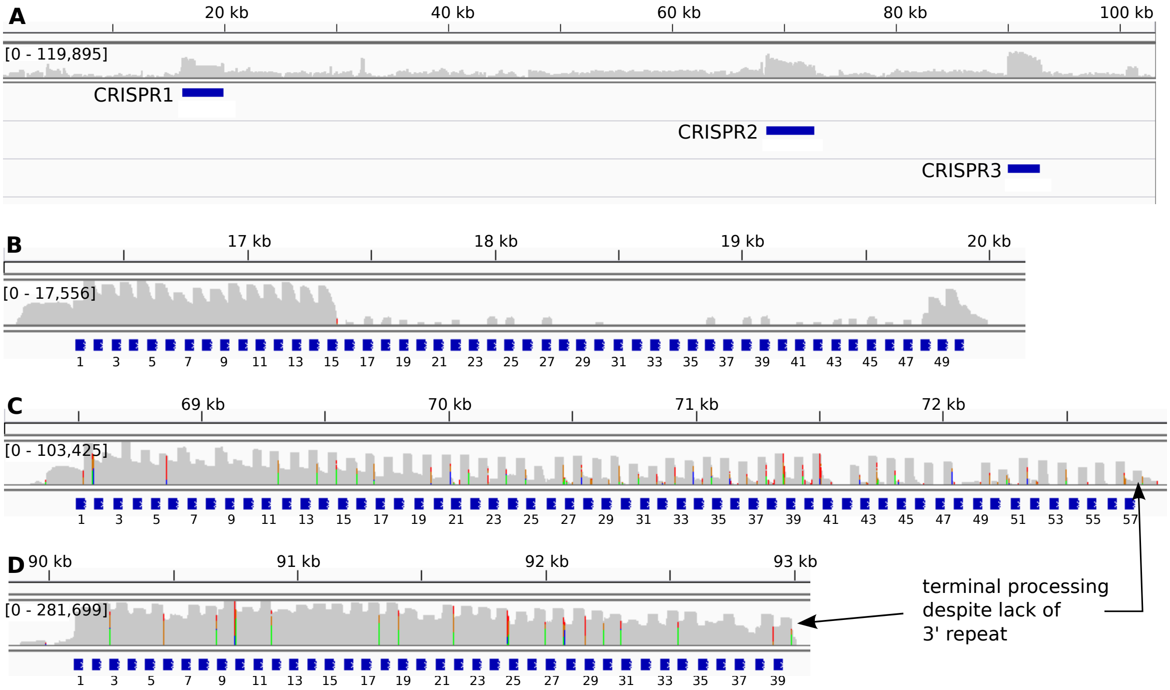


Figure 5.2. High expression levels of CRISPR-derived RNA on the pSYSA plasmid. (A) depicts the read coverage in log-scale (grey track) across the entire plasmid and the locations of the CRISPR1–3 are annotated as blue bars. All three CRISPRs are the most abundantly expressed loci on the plasmid. (B–D) show the expression profiles for CRISPR1–3, respectively. The reads have been filtered to reduce noise and the grey tracks in (B–D) depict their coverage profiles in log-scale. The numbers in the square brackets represent the absolute read number range; CRISPR3 is clearly the most abundantly expressed in comparison with the other two. The repeats are marked below by blue squares with their occurrence number. Due to the consecutive duplications of repeat-spacer units in CRISPR1 and CRISPR2, a unique mapping was impossible for these spacers so that their coverage appears identical. Moreover, CRISPR2 and CRISPR3 show a terminal processing despite the fact that there is no downstream repeat. Figure taken from [P10].

lengths 213, 124, and 1 nt, respectively [P10]. It is unknown whether the fact that CRISPR3 is basically leaderless (with only 1 nt) could affect new spacer acquisition; however, the array is clearly processed (Figure 5.2).

We observed an extremely high level of CRISPR-derived RNA transcripts, especially in comparison with other loci on the pSYSA plasmid (Figure 5.2.A). CRISPR3 RNA was most abundant with more than two million reads; almost 7-fold and 19-fold more than CRISPR2 and CRISPR1, respectively. Only very few reads (a total of 110, 60, and 1,430) mapped to the reverse strand; the majority mapped to the forward strand of the CRISPR1–3 arrays. This suggests only a very minor effect of technical bias introduced by the reverse transcription and sequence analysis.

To gain a more accurate picture of the CRISPR array expression, we filtered the original mapped `RNA-seq` reads (Figure 5.2.A) to reduce noise. The bulk of noise arose from short sequence reads that cover only the repeat regions and were therefore incorrectly mapped to all repeat instances, obscuring the coverage profiles. Thus, we selected reads that mapped with a read quality of 1, had an edit distance ≤ 2, were located on the forward strand, and had a unique match. Due to the duplications in CRISPR1–2, we also allowed reads for these loci that mapped to two locations. Note that the performed filtering delivered a clarified picture, but did not considerably change the original coverage profiles. In Figure 5.2.B–D, we present close-up profiles of the filtered read coverage for each of the CRISPR arrays. The CRISPR loci had a greater read coverage at the 5' end in comparison with the 3' end, which was also observed by Hale and colleagues [125]. Despite the generally high abundance of reads for all three CRISPRs, we noticed a lack of coverage corresponding to the repeat-spacer units 15–47 in CRISPR1 (Figure 5.2.B). This lack is due to a deletion of 33 repeat-spacer units in the recently sequenced substrain "PCC-M" [218] (used here) in contrast to the original plasmid sequence [158]. Consequently, only 16 crRNAs were expressed from the CRISPR1 locus in the "PCC-M" substrain.

5.4 Inferring processing characteristics from `RNA-seq` data

In agreement with their characterisation as distinct types of CRISPR-Cas systems, processing intermediates and mature crRNAs of different characteristic lengths were observed (Figures 5.3 and 5.4). We established cleavage sites and the boundaries of accumulating transcripts by counting the total number of 5' and 3' read starting and ending positions, relative to the closest direct repeat (summarised across all repeats across one array), using the `RNA-seq` dataset A (Figure 5.3). Note that due to the ligated poly(A) tails in the `RNA-seq` protocol, 3' read ends were not well defined for sequences ending in *A*'s, leading to staggered peaks. The repeat cleavage sites were most obvious with clear peaks of 5' read starts giving rise to the well-published 5' crRNA tags [32, 103, 129–131, 248, 260, 274, 326]. The 5' tags of CRISPR1 and CRISPR2 were identical (*ACUGAAAC*) and their length of 8 nt is in agreement with previous results. The 5' tag of CRISPR3 is unusual by having a length of 13 nt. Its sequence *AUUGAUUGGAAAC*, however, exhibits similarities to tags in many other published CRISPR repeats. In Part II, Chapter 3, we established that the 8 nt tag *AUUGAAAC*/*G* is conserved in 40 % of the 40 sequence families identified and other families contained minor deviations (Section 3.4.2). The similarities between this 13 nt tag in CRISPR3 and the aforementioned conserved 8 nt motif are clear: we merely observe a duplication of the prefix *AUUG* and of the *G* before finishing with the extremely common *GAAAC* suffix.

Concerning the number of observed cleavage events, CRISPR1 and CRISPR2 displayed only single cleavage sites within their repeats, whereas CRISPR3 was processed with a double cleavage activity. Interestingly, the first cleavage occurred at the 5' end of the repeats, mostly within the spacers. This result is supported by two observations (Figure 5.3): (1) 3' read ends in the spacers were immediately followed by 5' read starts, defining a clear cleavage site, which was not the case for the cleavage site at the 13 nt tag and (2) there is no accumulating RNA species that spans across the cleavage site in the spacer, whereas the 72 nt intermediate spans across the 13 nt cleavage site.

Figure 5.4 shows the read lengths that were accumulated in the `RNA-seq` data. To identify whether these lengths corresponded to intermediate and mature crRNAs, we calculated (from the filtered sets Figure 5.2.B–D) the percentage of reads that mapped to the locations

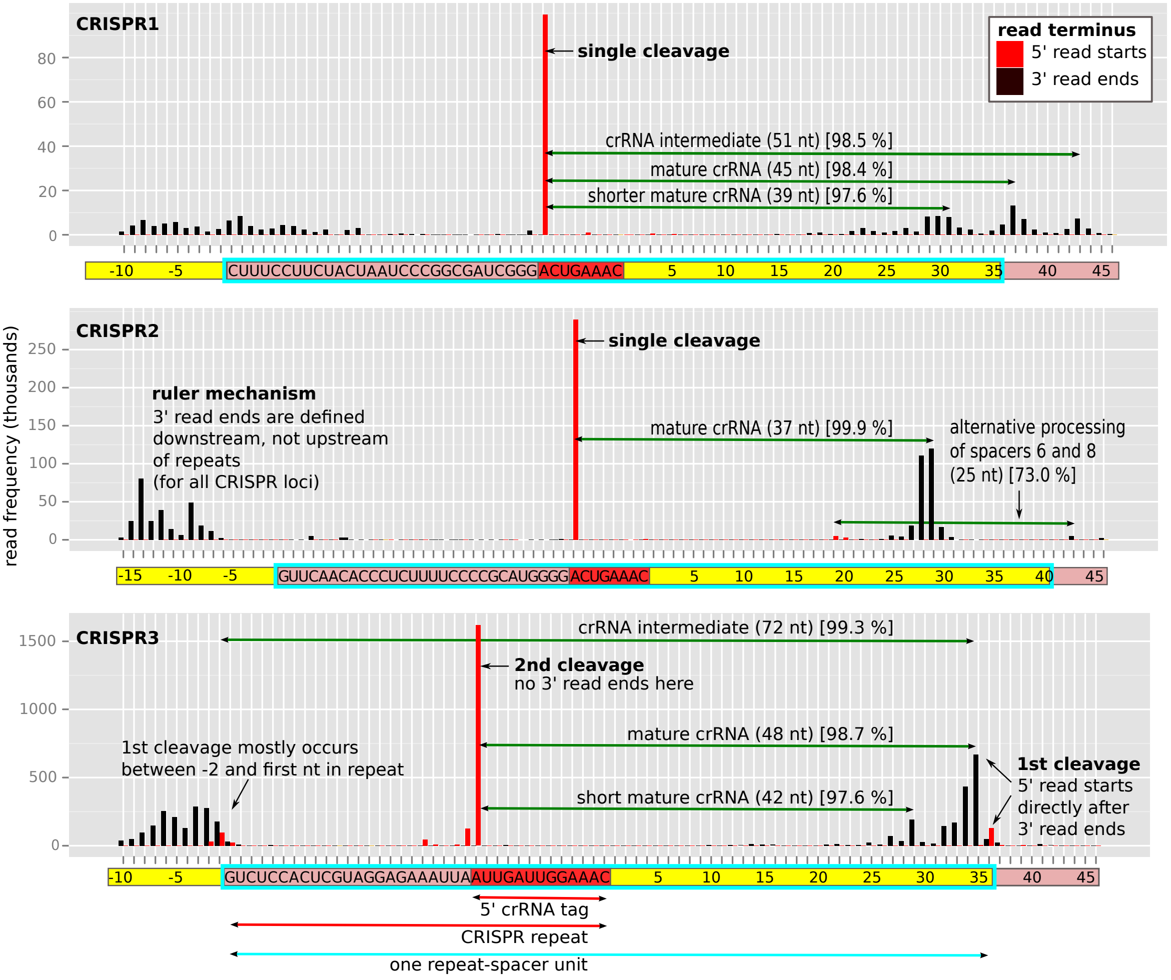


Figure 5.3. Frequency of read termini shows clear cleavage sites and distinct processing features. The number of reads (y-axis) starting (red) or ending (black) at a position relative to the closest repeat (x-axis) across an entire CRISPR locus illustrates the CRISPR maturation products (for `RNA-seq` dataset A). The repeat sequence is indicated in the pink+red, the 5' crRNA tag in the red, and the relative position in the spacer in the yellow rectangles, respectively (x-axis). One repeat-spacer unit is framed by the thick cyan square (due to different spacer lengths, the mode is illustrated). The green arrows correspond to the most abundant reads, i.e., the processed mature crRNAs or intermediate products. Albeit spacers of different lengths, we clearly see the ruler mechanism as the mature crRNA is trimmed to fixed lengths. We identified the location of the accumulating reads by giving the percentage of reads in the respective read-length category that map to the illustrated location (square brackets). For CRISPR3, the first cleavage site is in the spacer (not in the repeat), supported by two observations (1) reads only end at the cleavage site in the spacer, not in the repeat, (2) there is no accumulating RNA species that spans across the cleavage site in the spacer, whereas the 72 nt intermediate spans across the 13 nt cleavage site. CRISPR1–2 display only single cleavage sites and crRNAs are subsequently trimmed to their final length. CRISPR1 and CRISPR3 both have a second, less abundant mature crRNA transcript, which is exactly 6 nt shorter, whereas CRISPR2 only has one accumulating product. Note: Fluctuations of about 1–3 nt at read endings are due to the ligation of poly(A) tails in the `RNA-seq` protocol; real reads ending in *A*s cannot be determined correctly. Figure taken from [P10].

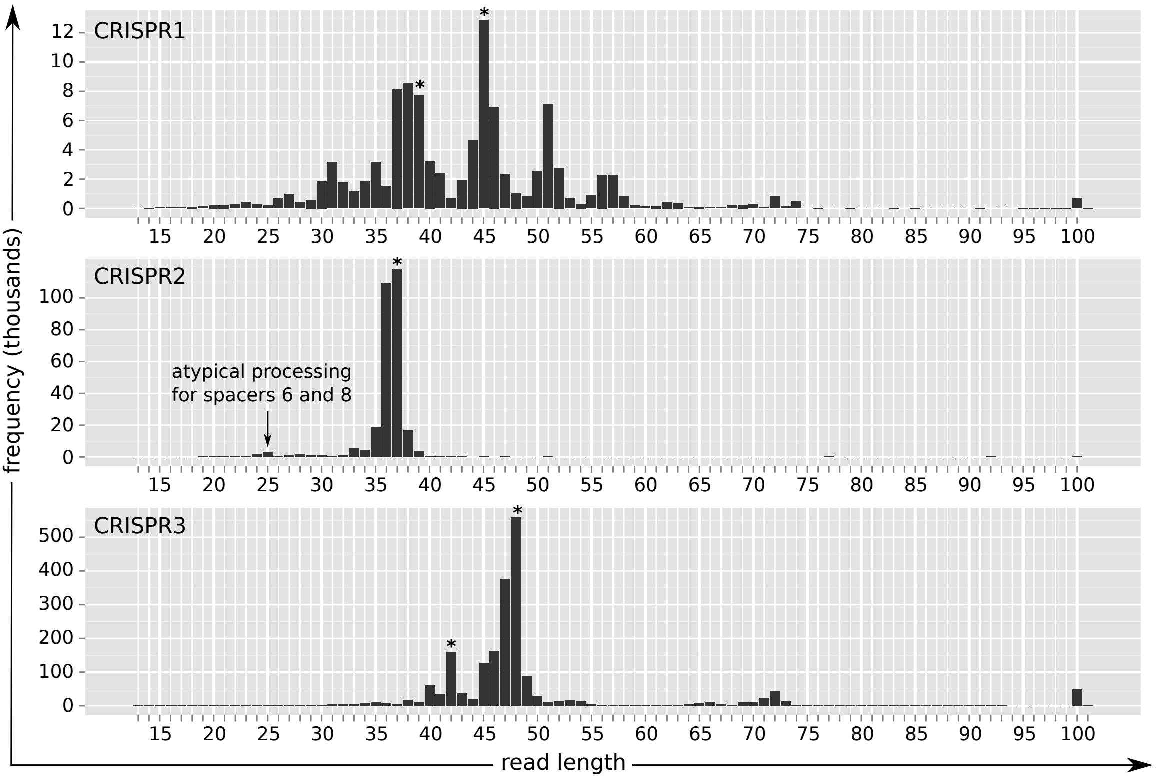


Figure 5.4. Accumulation of CRISPR RNA indicate lengths of mature crRNAs and intermediates. Read frequencies (y-axis) for all CRISPR loci, computed from `RNA-seq` dataset A. Read lengths are given on the x-axis, whereby it is important to note that the poly(A) tails of the `RNA-seq` protocol obscure read ends such that lengths of reads ending in A's cannot be determined exactly. The transcripts assumed to correspond to mature crRNAs are marked by stars. Figure taken from [P10].

indicated in Figure 5.3 out of all reads with the respective characteristic lengths (percentage in square brackets and 1–2 nt position-specific variation was allowed). The high percentages gave convincing evidence that the indicated locations are correct. The most probable mature crRNAs are 45 and 39 nt for CRISPR1, 37 nt for CRISPR2, and 48 and 42 nt for CRISPR3. Notice that for CRISPR1 and CRISPR3, two accumulating species of mature crRNAs existed, which were both 6 nt different in size, and the longer transcript was more abundant (both observations were previously seen in *Pyrococcus furiosus* [125] and *Staphylococcus epidermidis* RP62a [129]). Despite the common difference of 6 nt in mature crRNA lengths for CRISPR1 and CRISPR3, other distinct features existed: In all Northern hybridisations, double bands were observed for CRISPR3 [P10], which indicated two distinct lengths (6 nt apart) for each accumulating (intermediate) transcript species; whereas for CRISPR1 this was not observed. Instead CRISPR1 transcripts accumulated to multiple lengths, which were all shorter than one repeat-spacer unit (71 nt). This alludes to a final stepwise trimming of one repeat-spacer unit subsequent to the cleavage at the 8-nt tag in both adjacent repeats.

Albeit varying lengths of the spacers, crRNAs for all loci accumulated to fixed characteristic lengths (Figure 5.3), which further supports the ruler mechanism published for the Csm and Cmr systems [125, 129]. Moreover, although the final repeat was cleaved at the usual position for all loci, only CRISPR2–3 displayed a notable accumulation of a 3' terminal transcript downstream from the last repeat (Figure 5.2.C–D). These terminal transcripts were of equal length with their respective mature crRNAs, albeit no second 3' repeat sequence; not even a partial, or a mutated repeat sequence could be detected. These terminal, potential crRNAs indicate that the 5' repeat is the anchor of the ruler mechanism and that this measured crRNA accumulation is independent of a subsequent cleavage in the downstream repeat. This was not observed for CRISPR1, which further supports the previously mentioned, step-wise 3' trimming.

In summary, while the described processing patterns shared previously published common features, detailed evidence suggests distinctly different pathways. Most importantly, CRISPR1 displays a 6-nt step-wise trimming of the 3' end of all accumulating RNA species, whereas, for CRISPR1–2, intermediate RNA species display a final cleavage, measured from the 8-nt tag at the 5' of the intermediate RNA, to produce the mature crRNA.

5.5 Stability of crRNAs may be dependent on spacer structure

We observed vast differences in the processed crRNA abundances across the CRISPR arrays (note that the log-scale reduces the visible differences in Figure 5.2). Given that each CRISPR array has only one TSS and is thus transcribed as one transcript, no obvious reason for major differences in accumulation exists. This variability could be partially explained by the stability of the crRNA-Cas protein complexes: highly structured crRNA could obstruct their formation, leading to crRNA degradation. To test this idea, we compared the ratio of degraded products to full-length crRNA with different structural properties of the CRISPR array.

5.5.1 Calculation of crRNA degradation

We estimated the degradation rate of crRNAs from the `RNA-seq` data by the ratio of the number of reads that were substantially shorter than a full-length mature crRNA to the number of reads that represent a full-length crRNA at a single repeat-spacer unit location in the CRISPR array. Let i_s be the starting and e_s be the ending position (in the genome) of the current spacer s, and i_r be the starting and e_r be the ending position of the current read r. We then considered all reads starting with $i_r > i_s - 25$ and $e_r < e_s + 10$ to represent processed full-length crRNA sequences, called read set C. Of these reads C, we selected the possibly degraded reads (set D) with $i_r > i_s - 8$ and $e_r < e_s - 10$ (we used $e_r < e_s - 15$ for dataset A, because very many reads seemed stable between $e_s - 10$ and $e_s - 14$). It is difficult to select this 3' cutoff because it is unknown until which length the crRNA is still functional, i.e., can locate its target. The 5' cutoff is easier due to the fixed cleavage site at $e_s - 13$ (for CRISPR3). The number of potentially degraded crRNA was then normalised by the total number of reads at that crRNA locus to obtain a degradation ratio: $degradation\ ratio = D/C$. We compared several properties of the mature crRNA to this ratio to determine possible factors for higher degradation rates.

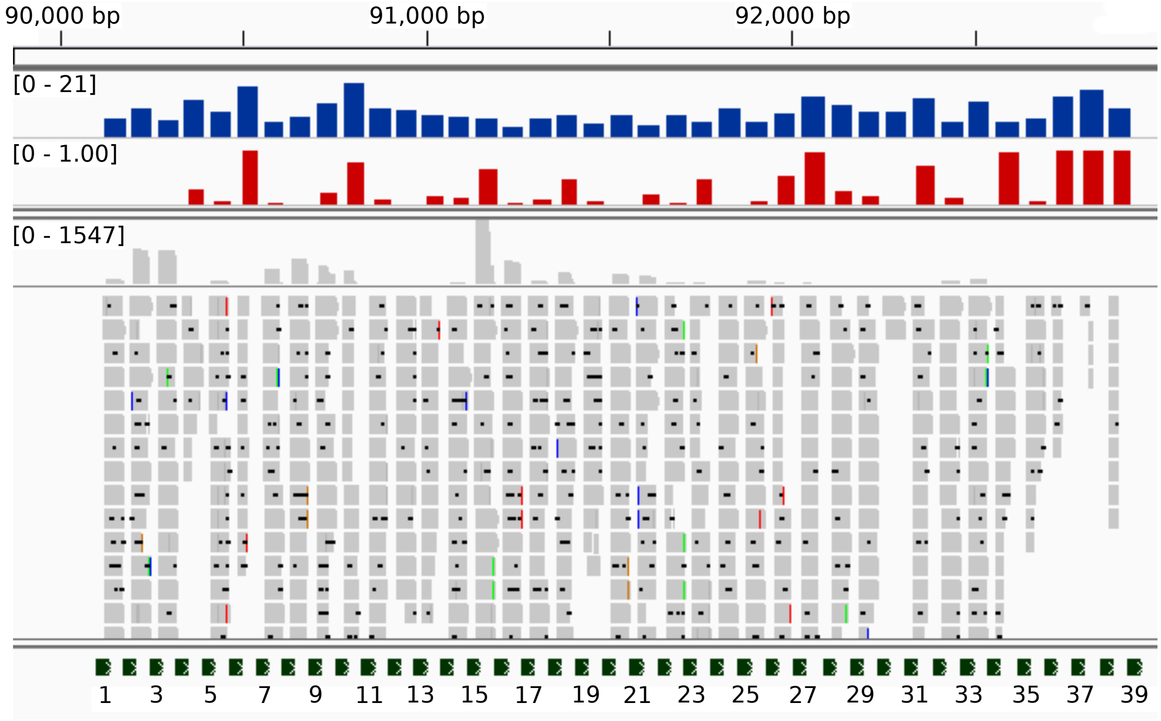


Figure 5.5. A significant relationship was measured between the 'structuredness' and the assumed degradation rate of individual crRNAs. The degradation of mature crRNAs correlates with spacer ensemble energies with a Pearson's correlation coefficient $r = 0.56$ and $p = 0.00025$ (`RNA-seq` dataset B). Depicted is the CRISPR3 locus on the pSYSA plasmid of *Synechocystis* sp. PCC 6803 with the following tracks: (blue) The absolute ensemble energy of the spacer sequence as determined by `RNAfold` (greater values correspond to more stable structures); (red) the normalized degradation profile of previously processed crRNA; (grey) sequence reads corresponding to degraded or full-length mature crRNA; (green) the CRISPR-repeat locations. Some crRNA positions remain full-length, whereas other positions are degraded (grey track). We selected only reads that correspond to mature crRNAs. Reads that cover two spacers were excluded for this analysis since they correspond to crRNA precursors. Figure taken from [P10].

5.5.2 The relationship between spacer 'structuredness' and crRNA degradation

The most convincing correlation between degradation and RNA structure was seen in the ensemble energy of the separate spacer sequences (Figure 5.5, blue track) with a Pearson's correlation coefficient of 0.56 ($p = 0.00025$) for CRISPR3. High ensemble energies correspond to spacers that can form more stable secondary structures. This indicated a strong relationship between the "structuredness" of the spacer and the degradation ratio of previously processed crRNA: more stable structures could lead to a higher rate of degradation (note that we give the absolute ensemble energy values and that in reality a negative correlation exists, due to negative energies). More precisely, all spacers in the CRISPR3 array with an ensemble energy below −15 kcal/mol had the highest degradation ratios. This result was also achieved for the smaller `RNA-seq` dataset B. Albeit the statistically significant correlation for the larger dataset A at $r = 0.38$ and $p = 0.018$, the correlation in this set is not as convincing, which is likely due to the differences in the datasets: In dataset A, only about 4 % of the reads were short enough to be considered as degradation products. It is unlikely that the signal was strong enough to be detected in this minor subset of reads, whereas in dataset B, the ratio of possible degradation products in comparison with non-degraded reads was much higher (see grey track in Figure 5.5). CRISPR1 and CRISPR2 could not be analysed for correlation to structuredness because too few reads mapped to these loci in dataset B.

In spite of transcripts arising from a single TSS, mature crRNAs accumulated to significantly different abundances implying differences in their stabilities. Our computational analysis of CRISPR3-transcript accumulation indicated that spacers forming more stable structures are linked to higher degradation rates of the crRNA sequence. A similar observation has recently been reported for the crRNAs derived from CRISPR locus C in *Sulfolobus solfataricus*, where those crRNAs with the potential to fold into the more stable structures were clearly less abundant than those with only modest folding propensity [354]. Interestingly, the studied *S. solfataricus* system is of CRISPR subtype III-B, similar to the CRISPR3 of *Synechocystis* studied here. Thus, the different quantities of mature crRNAs could be due to their different loading efficiencies into the CMR complex. A highly structured crRNA could prevent or delay the RNP (ribonucleoprotein) complex formation and thus lead to a lack of protection and consequently higher rates of degradation. Therefore, the more efficient spacer is likely one that remains mostly unstructured.

5.6 Conclusion

The cyanobacterium *Synechocystis* sp. PCC6803 harbours three distinct CRISPR-Cas systems, CRISPR1–3, on a single plasmid. Analysing `RNA-seq` data for the CRISPR1–3 loci, we found that transcripts from all CRISPR arrays were highly abundant, especially in comparison with other loci on the pSYSA plasmid. Notably, the individual crRNAs had profoundly varying abundances despite single transcription start sites for each array. A more detailed analysis determined the length and locations of accumulating intermediate and mature crRNA species. In addition, the most frequent 5'- and 3'-read-end mapping locations gave a detailed insight into cleavage sites and processing patterns and especially highlighted the fact that the crRNAs from each locus must have been generated by distinct pathways. In a final analysis, CRISPR3 spacers with stable secondary structures displayed a greater ratio of degradation products. These structures might interfere with the loading of the crRNAs into RNP complexes, explaining the varying abundances.

In conclusion, the analysis of `RNA-seq` is an appropriate method for not only establishing general CRISPR RNA abundances but can also be used to determine detailed processing signals and to characterise accumulating RNA species.

Part IV

Structure prediction in long RNAs

Part IV: Structure prediction in long RNAs

Life offers us both problems and solutions. It is for us to choose what we want.

The regulatory function of RNA largely depends on its structural conformation in addition to sequence-specific binding affinities. Most analyses of RNA structure focus on regulatory ncRNAs as these are usually reasonably short with globally conserved structures [7,108,111]; notable examples are transfer RNAs, ribosomal RNAs, and miRNAs. RNA regulatory function is not only guided by such global structures, but can be influenced by local RNA structures that only form in a subsequence of a long RNA, e.g., mRNA, lncRNA, or precursor ncRNA. In the literature, local mRNA structure—in particular whether the local region of the structure is single stranded—has been considered important for the binding of *trans* factors, such as RBPs and small ncRNAs [109,143,164,171,234,258,303]. In addition, structured *cis*-regulatory elements[1], frequently located within untranslated regions (UTRs) of mRNAs, are involved in regulating the mRNA they are located within [149]. The structure of an mRNA at a binding site of a *trans* factor or which forms a *cis*-regulatory element is *local* in the sense that it only involves a small subsequence of the full mRNA sequence, and the *global* structure of the entire mRNA is irrelevant to the regulatory function. In contrast to small ncRNA structures, little research has been dedicated to the more challenging task of elucidating the structural properties of long RNA, e.g., mRNA, lncRNA, or precursor ncRNA.

The main goal in Chapter 6 is to provide a guideline on how to *best* elucidate local structure properties of long RNA molecules when this is required for the analysis of individual sequences. Although the emphasis here is on mRNA structure, results should be applicable to other long RNA species. In fact, the knowledge gained in Chapter 6 was applied to the prediction of regulatory structure motifs in CRISPR arrays in Chapter 7. In addition, we used the CRISPR array as a platform to explore the effect of surrounding sequence context on motif-structure formation and demonstrated how unfavourable sequence contexts can abolish biological function. Results presented in this part were published in [P4,P7,P10], or are currently under review [P8].

[1] See Figure 2.2 in Section 2.1.1 for some examples of *cis*-regulatory structure elements.

CHAPTER 6

Predicting secondary structure in mRNAs

Regulatory elements are located predominantly in the 3'UTR of an mRNA. These can either be simple binding motifs of about 6 to 20 nt or more complex structured *cis*-regulatory elements that can involve up to a few hundred nucleotides [264]. Not only the stable formation of base pairs involved in structured *cis* regulatory elements—but also the structural accessibility of simple binding motifs—is generally important for the regulatory function [109, 143, 164, 171, 234, 258, 303]. For computational detection and characterisation of regulatory elements on mRNAs, we require methods for accurate structure prediction. There are two aspects to consider when analysing mRNA structure: (1) the structure influencing the function of regulatory elements is *local*, i.e. involves only a subsequence of the mRNA, and (2) mRNAs are usually hundreds to thousands of nucleotides long. In this work, we only consider approaches that compute binding scores for all possible base pairs, e.g. base pair probabilities. These structure-prediction approaches are either global or local (see Section 2.5). A global structure prediction is one where all possible base pairs for the entire input sequence are considered, and a local structure prediction is one where base pairs with long spans (see Definition 2.8 in Section 2.4.1) are ignored.

As most algorithms for structure prediction of RNA have been developed for ncRNAs, the first approach to predicting structure in mRNAs (or long RNAs in general) would be to use either `UNAfold` [206], `RNAfold` [138] or `RNAstructure` [255] on the entire mRNA sequence or at least the entire 3'UTR. The cubic time and space complexity of these *global* approaches, required to determine probabilities for all possible base pairs within a sequence, makes their application to very long RNAs infeasible. The most basic solution to long runtimes would be to extract the part of interest from the mRNA and fold this globally. However, as shown later, this approach can lead to border effects at the artificial sequence ends and it also ignores the influence of base pairs directly adjacent to the window[1]. A first algorithmic solution to

[1] The consequence of ignoring the context sequence around a region of interest is explored further in Chapter 7.

the high runtime complexity of global structure prediction was to limit the distance on the sequence between two base pairs, i.e. the base-pair span (Definition 2.8, Part I) and to ignore any base pairs with spans larger than a given threshold, typically denoted by L. As this approach still folds the entire input sequence simultaneously and merely restricts the base-pair spans of the predicted structures, we considered it to be *semi-local*; implementations are `RNALfold` [139] to find locally stable structures, `Rfold` [170] for base-pair probabilities, and `Raccess` [171] for accessibilities. The second algorithmic solution was to predict structures in sliding windows of a fixed length denoted by W, in addition to the maximum base-pair span constraint L [18, 19]. The likelihood of a base pair occurring is now an average of all base-pair probabilities for all windows in which it occurs. This *window-based* approach is *local* in the sense that each window is folded independently of the rest of the sequence. Nevertheless, a single window is folded semi-locally as before. Approaches that predict true local structures, without the use of fixed windows, were not available. The window-based approach is implemented in `RNAplfold` [18, 19]. RNA structure prediction algorithms are introduced in more detail in Section 2.5.

Although `RNAplfold` is currently the most popular tool for elucidating secondary structure in long RNAs, especially for calculating accessibility of potential target sites, e.g. [161, 194, 205, 303], reliable benchmarks of the accuracy of various tools and appropriate parameter settings have not been performed[1]. It has been shown that the majority of base pairs have short spans [74, 233]; therefore, it can be assumed that local approaches give an accurate approximation of structure. However, the impact of long-distance base pairs and surrounding context structure—and, in fact, the performance of *local* in comparison with *global* approaches—had not been quantified before this work.

Previous investigations of the locality parameters W and L were centred around specific applications. For example, Tafer *et al.* evaluated effects of accessibility on the efficacy of small interfering RNA interactions [303]. Folding parameters that achieved the most significant results, a window size of $W = 80$ nt and a maximum base-pair span of $L = 40$ nt, were subsequently used as standard values for local secondary structure predictions [161, 194, 205]. Similar analyses were performed in [171, 288]. A window size that was equal to the maximum base-pair span was used in [170] and it is also the default setting in `RNAplfold`. Our subsequent benchmark analysis showed that these previously used parameters perform poorly. For the first time, we showed that a local approach is not only more practical, but that it even outperforms its global counterpart in accuracy when predicting secondary structure in mRNAs.

6.1 LocalFold: reducing window-border effects

In the standard window-based approach, used for `RNAplfold`, base-pair probabilities are computed for each window separately and then averaged over all windows that the respective

[1] Initially, local approaches were applied due to their practical runtimes.

base-pair occurs in (see Equation 2.4 in Section 2.5.4). To address potential prediction biases at window borders, we developed a modified version of the standard window-based approach that ignores these predictions, which we called `LocalFold`. For the purpose of describing the modification for `LocalFold`, we rewrite the equation for `RNAplfold` as:

$$p_{avg}^{L,W}(i,j) = \frac{1}{|\mathbb{W}(i,j)|} \cdot \sum_{\mathcal{W}^u \in \mathbb{W}(i,j)} p^{\mathcal{W}^u,L}(i,j), \tag{6.1}$$

where W is the window size, L the maximum base-pair span, $\mathcal{W}^u$ is the window beginning at position u, and $\mathbb{W}(i,j)$ is the set of all windows that include the base pair (i,j). For `LocalFold`, we modified the calculation, such that $\mathbb{W}(i,j)$ contains only windows where either bases r_i and r_j were not within the first or last b positions of the window. Window borders that coincide with the input sequence ends are exempt from the modification and are calculated as in `RNAplfold`.

The `LocalFold` algorithm is applicable to all parameter combinations of W, L, and b satisfying $W - L \geq 2b$. The method is thus limited to a W that is sufficiently larger than L. The b parameter does not exclude any parts of the sequence; the filtering induced by b merely ignores the outliers in the averaging calculation (Equation 6.1). The parameters are set to $W = 200$, $L = 150$, $b = 10$ by default. We recommend to use $b = 10$, since this achieved the best result and clearly eliminated most of the bias at the borders (Figure 6.3). The time and space complexity stays the same as for `RNAplfold` [18, 19]. `LocalFold` is available for download at `www.bioinf.uni-freiburg.de/Software/LocalFold/`.

6.2 Evaluating the stability of local structure motifs

When comparing the performance of structure-prediction approaches, we should generally have a set of *true* RNA structures for their respective sequences. Then, a comparison of prediction results depends on the measurement of the predicted stability of these true structures. This comparison is, however, complicated by: (1) the approaches compute either probabilities[1] or averaged probabilities[2] for individual base pairs; (2) we require a measure of stability for complete structures and not only single base pairs; and (3) to our knowledge, a measure to compare structure stabilities computed by either global or local prediction approaches has not been addressed in the literature prior to this work. More precisely, in the investigation of *cis*-regulatory elements, we required a measurement for the stability of a local structured element within a greater context. Therefore, we needed to determine the accuracy of the prediction of the entire element based on individual base-pair scores. In the literature, there was no consistent measure for this purpose, however, structure stability measures have been applied to global structures [37, 72, 197]. We generalised the previously applied measure of structure accuracy to local structure prediction.

[1] `RNAfold`, `Rfold`, `Raccess` compute base-pair probabilities.

[2] `RNAplfoldand` `LocalFold` compute average base pair probabilities for all windows the base pairs occur in.

Let R be an RNA sequence, and S_l be a local structured element in R. The accuracy $\mathcal{A}$ is the expected overlap of a local structure S_l and a global structure S of R:

$$\begin{aligned}\mathcal{A}(S_l|R) &= \sum_{S\in\mathcal{Q}_R} |S_l \cap S| \cdot Pr[S|R] \\ &= \sum_{S\in\mathcal{Q}_R}\sum_{(i,j)\in S_l} \mathbf{1}\{(i,j)\in S\}Pr[S|R] \\ &= \sum_{(i,j)\in S_l}\sum_{S\in\mathcal{Q}_R} \mathbf{1}\{(i,j)\in S\}Pr[S|R] \\ &= \sum_{(i,j)\in S_l} p(i,j). \end{aligned} \tag{6.2}$$

$\mathcal{Q}_R$ is the ensemble of all possible structures for R; $\mathbf{1}\{(i,j)\in S\}$ is an indicator function that is 1 if $(i,j)\in S$ and 0 otherwise; the probability of observing a structure, $Pr[S|R]$, is given in Equation 2.1, Section 2.5.2. In simple terms, the accuracy of a local structure is the sum of all its base-pair probabilities in the global structure ensemble.

For window-based approaches, the probability of observing a given base pair (or structure) in a window is comprised of the probability for choosing the window $\mathcal{W}^u$ (beginning at position u) and the probability of observing the base pair (structure) in $\mathcal{W}^u$. Each window has an equal probability and the structures within each window are Boltzmann distributed as in global folding [210]. Thus, to gain single scores per base pair (i,j), `RNAplfold` averages over all windows containing the base pair, $\mathbb{W}(i,j)$ (Equation 6.1).

Regarding the accuracy of a local structure element S_l, we define $\mathbb{W}(S_l)$ to be the set of windows that contain the complete structure S_l, similar to the definition in the case of a base pair (see Equation 6.1). Then we define the average accuracy as:

$$\begin{aligned}\mathcal{A}_{avg}(S_l) &= \frac{1}{|\mathbb{W}(S_l)|}\sum_{\mathcal{W}^u\in\mathbb{W}(S_l)} \mathcal{A}(S_l|\mathcal{W}^u) \\ &= \frac{1}{|\mathbb{W}(S_l)|}\sum_{\mathcal{W}^u\in\mathbb{W}(S_l)}\sum_{(i,j)\in S_l} p^{\mathcal{W}^u,L}(i,j).\end{aligned}$$

If we had the same windows for each base pair in S_l, i.e., for all $(i,j)\in S_l, \mathbb{W}(i,j) = \mathbb{W}(S_l)$, where $\mathbb{W}(i,j)$ is the set of windows that contain the base-pair (i,j), then analogously to Equation 6.2, we could continue with

$$\begin{aligned}\mathcal{A}_{avg}(S_l) &= \sum_{(i,j)\in S_l}\frac{1}{|\mathbb{W}(i,j)|}\sum_{\mathcal{W}^u\in\mathbb{W}(i,j)} p^{\mathcal{W}^u,L}(i,j) \\ &= \sum_{(i,j)\in S_l} p_{avg}^{\mathcal{W}^u,L}(i,j).\end{aligned} \tag{6.3}$$

Having the same set of windows for each base pair, however, could only be enforced if the location of the element was known in advance. Since this is not the case when searching for local structures, we used Equation 6.3 as an approximation of the average accuracy of the local structure S_l.

For the comparison of accuracies for structure elements of different sizes, we normalised them by the number of base pairs within the respective local structure S_l:

$$bp\text{-}accuracy(S_l) = \frac{\mathcal{A}_{avg}(S_l)}{|S_l|}, \tag{6.4}$$

and analogously, we substituted $\mathcal{A}_{avg}(S_l)$ with $\mathcal{A}(S_l)$ for the non-averaged base-pair probabilities.

Intuitively, the *bp-accuracy* is the mean base-pair probability of all base pairs within the reference structure (i.e. *cis*-regulatory element); it measures the thermodynamic stability of the structure within its global context. The *bp-accuracy*, however, does not consider false positive base-pair predictions. No gold standard for negative base pairing exists and it was unclear when a base pair that is not part of the local structure should be regarded as negative, or incorrect. For example, one could consider all possible conflicting base pairs, i.e., all base pairs involving one and only one base from a correct base pair, to be incorrect (in a secondary structure, a base can only be paired to one other). This is problematic for three reasons: (1) there are about $2L$ more incorrect than correct base pairs; (2) a different number of negative base pairs would occur for different L values, hence, it is difficult to compare global and local folding methods; and (3) it is unknown to what extent the mRNA folds into different conformations, or refolds. Alternative structures do exist *in vivo*, e.g. in riboswitches [28]; some conflicting base pairs could be true variants. Kiryu *et al.* proposed a way to calculate specificity by considering all base pairs predicted in random sequences to be incorrect [170]. Randomly designed RNA sequences, however, could also form stable structures [263].

In conclusion, we used the *bp-accuracy* to compare the stabilities of given local structure motifs within base-pair predictions calculated by both global and local structure-prediction approaches.

6.3 `CisReg`: a curated set of *cis*-regulatory elements

Having established general approaches to predicting mRNA structure and a measure to compare predicted structure stabilities, we required a suitable dataset of local RNA structures. An important benchmark of new mRNA structure discovery methods is their ability to accurately predict known *cis*-regulatory elements. These known elements are characterised in several databases, of which the largest is the RNA families database (`Rfam`) [95, 111]. Release 10.0 contained 1,446 covariance models, mostly for non-coding RNA genes, but also for structured mRNA elements [95]. Each model consists of a set of published “Seed” and computationally extended “Full” alignments. Sequences within the structure alignments consist of only the structured element, and usually lack the flanking sequence from the mRNA, needed to assess structure prediction.

For this study, we curated a new benchmark set for mRNA *cis*-regulatory elements. We extracted and individually re-examined a set of 95 families of *cis*-regulatory elements from

Rfam that were correctly classified and adopted secondary structures without pseudoknots[1]. Of these, 24 were from eukaryotic mRNAs and 71 from prokaryotic or viral genomes. The eukaryotic mRNA elements had diverse functions (e.g., mRNA localisation, translation efficiency or mRNA stability) and most were located within 3'UTRs. A large number of the genomic elements were from RNA viral genomes or from bacterial mRNAs. For each element, we extracted three different lengths of flanking regions from the mRNAs (including coding regions and 5'UTRs), or from the genomes when these were not available: 100, 200, and 500 nt, or otherwise to the sequence ends. Subsequently, we filtered and processed the elements to maximise structure integrity and a small proportion of sequences were excluded as they did not match sequences in the EMBL Nucleotide Sequence Database. The exact data preparation process and a redundancy analysis are provided in Section D.3.

The CisReg dataset used in this study consists of 2,500 individual elements (95 families) with over 85,000 base pairs, and we propose it as a reference set to test future prediction algorithms. Furthermore, we provide a website for the data including additional information and statistics: http://lancelot.otago.ac.nz/CisRegRNA/.

6.4 Benchmarking preliminaries

To be able to determine the best approach for predicting secondary structure in mRNAs, we required suitable algorithms, large and high-quality datasets, and performance measures.

We made a careful selection of algorithms that reflect the current status of *secondary* structure prediction with a particular emphasis on local methods. Due to their broad usage, we concentrated on partition-function–based approaches that produce probabilities or average probabilities for base pairs, given an RNA sequence (see Table 6.1). The rationale behind comparing these algorithms is given in the introduction to this chapter and further details are described in Section 2.5. Execution calls are given in Section D.3.

The performance of LocalFold and current methods available for folding mRNA sequences was compared using two sets of data. First, the previously described CisReg data (see Section 6.3), containing >85,000 base pairs from 2,500 *cis*-regulatory elements, which were extracted from 95 hand-selected families from the Rfam database [95, 111]. Second, for the evaluation of the accessibility predictions we used the set of *in-vitro* secondary structure profiles from [165]. This set, referenced to as YeastUnpaired, consists of nucleotide-wise measurements for 3,196 mRNAs from *Saccharomyces cervisiae.* These profiles were derived by parallel analysis of RNA structure (PARS). With PARS the single-strandedness (as well as double-strandedness) of a set of sequences is inferred using a combination of RNase digestion and deep sequencing [165]. Kertesz *et al.* report that they covered approximately 100-fold more transcribed bases than all previously published footprints combined, making this dataset uniquely suited for a comprehensive analysis of prediction performance.

[1] Structures containing pseudoknots were ommited because these cannot be predicted by structure-prediction approaches, which are efficient enough to be applicable to long RNA sequences.

Table 6.1. Summary of the prediction methods and the benchmark datasets used in this work. L is the maximum base-pair span, W is the window size and b is the border size within which to ignore base pairs of a single window (see Sections 2.5 and 6.1). Table taken from [P4].

Method	Parameters	Type	Output
`RNAfold`	–	Global	Base-pair probabilities
`Rfold`	L	Local	Base-pair probabilities
`Raccess`	L	Local	Accessibilities
`RNAplfold`*	L, W	Local	Average base-pair probabilities and accessibilities
`LocalFold`*	L, W, b	Local	Average base-pair probabilities and accessibilities

Dataset	Description
`CisReg`	2,500 *cis*-regulatory elements in 95 `Rfam` families, filtered and processed in this work
`YeastUnpaired`	Data on the single-strandedness of single positions for 3,196 *Saccharomyces cervisiae* mRNAs from [165]

*Window-based approach

We previously defined and introduced the *bp-accuracy* (Equation 6.4) as a suitable measure to compare predicted base pair probabilities (Section 6.1) for local structure elements in the `CisReg` dataset. In the case of accessibility predictions, we compared the methods according to their ability to correctly classify paired and unpaired bases. Classification performance was measured using the AUROC (Section 2.3). This measure is independent of the types of outputs of the different algorithms. The accessibility of a base is the complement of the sum of all base-pairing probabilities that involve that base (see Equation 2.5), thus implicitly, the base-pairing distribution is taken into account. Therefore, the performance comparisons of accessibility should indicate which method produces the more accurate base-pair distributions.

Finally, we developed suitable tests that were designed to: (1) identify and elucidate the optimal degree of locality and (2) investigate the effects of artificial window borders and sizes, and (3) quantify the performance of each method on the two benchmark datasets. Prediction methods and datasets are summarised in Table 6.1.

6.5 How local is local structure?

The main difference between global and local prediction is the restriction of the base-pair span (*bp-span*, Definition 2.8) to a maximum of L. In this section, we explored how different settings for L affects prediction results, and identified how local *cis*-regulatory structures are (in general) so that a suitable parameter setting for L can be an informed decision, rather than a random choice. Further analyses were performed to clarify why local structure prediction is so accurate—even for relatively short maximum *bp-span* settings.

6.5.1 Best performance for a maximum *bp-span* between 100–150 nt

For local folding approaches, the main question was which degree of locality to use. To address this question, we compared `Rfold` predictions with L between 40 and 400 nt to (the global) RNAfold results using the CisReg data. Local folding was represented by Rfold because the introduction of the base-pair restriction is the only conceptual difference to global folding; whereas the window-based approaches introduced the window size (W) as an additional parameter. The lowest median *bp-accuracy* of 0.46 was achieved using Rfold with $L = 40$ (Figure 6.1.A). The accuracy increased with greater L values until a maximum of 0.59 was achieved at $L = 150$, after which accuracies decreased slightly. Rfold outperformed RNAfold at $L \geq 60$. The difference between the *bp-accuracy* distributions of Rfold ($L = 150$) and RNAfold was significant with $p = 1.2 \times 10^{-7}$ (two-sample Wilcoxon Rank Sum Test). The *cis*-regulatory structures in Figure 6.1.A were situated within a context of up to 500 nt to either side, the folded RNA sequence was thus only approx. $1,000$ nt long and often not the full-length mRNA. Therefore, we compared Rfold ($L = 150$) to RNAfold on the 179 available full-length mRNA sequences (Figure 6.1.B). Here the median base-pair accuracy of both methods was reduced, but the difference between the two methods increased: 0.13 compared to 0.07 in Figure 6.1.A.

When investigating the degree of locality L suitable for the YeastUnpaired data, we observed results similar to the CisReg data, see Figure 6.8 (the main discussion of this figure follows later). For accessibility, Rfold outperformed RNAfold at $L \geq 50$ and the performance increased up to the optimum at $L = 100$. $L > 100$ exhibited only a minor decrease in AUROC, thus L was robust to larger L values. Nevertheless, the quality of predictions decreases down to the level of RNAfold for both datasets: the greater the span L, the more global the prediction becomes until it *is* global when L equals the sequence length.

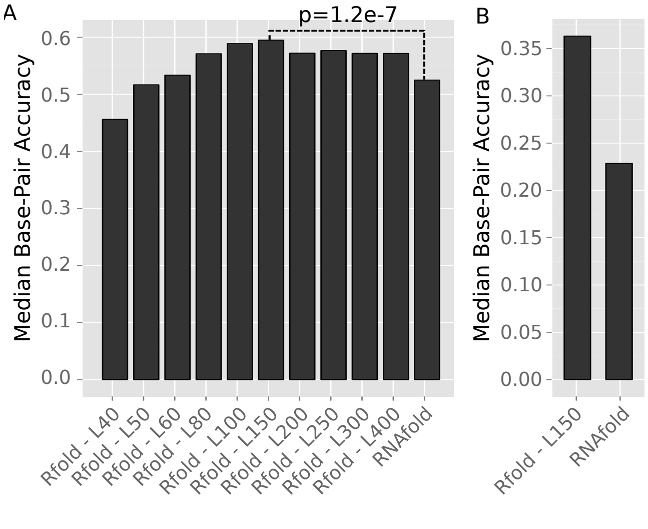


Figure 6.1. Comparison of global vs. local folding using the methods RNAfold and Rfold. The median base-pair accuracy (y-axis) is given for the CisReg dataset. (A) Comparison of RNAfold and Rfold using different L values. (B) A subset of the CisReg dataset that comprised of 179 full-length mRNA. Figure taken from [P4].

6.5.2 Most base-pairs have short spans

Our results on the best value for L reflected the distribution of base-pair spans within known structures [74, 233]: we observed that 83 % of all base pairs had a *bp-span* less than 100 nt (85 % $\leq$ 150) for all the *cis*-regulatory elements in the CisReg dataset (Figure 6.2.A). Thereafter, the increase in the number of base pairs with a larger span is very slow. Although we specifically chose local regulatory structures located on the mRNA, the distribution was similar to previously published data: Doshi and colleagues showed the same exponential decrease in observed base pairs with respect to increasing *bp-span* length in 496 16S rRNAs, with 75 % of all base pairs with *bp-span* $\leq$ 100 nt [74]. In 151 ncRNA structures from 151 seed alignments in Rfam, 85 % of the base pairs had a *bp-span* $\leq$ 100 nt [171]. The latter two analyses looked at global structures that form long-range base pairs. This observed exponential decrease in observed base pairs with increasing *bp-span*s implies that the majority of base pairs have short spans, i.e. are *local*; therefore, smaller L values ($L \leq 100$) performed comparably well. Because of the agreement of our results with the general distribution of base-pair spans, we suggest that local folding with restricted base-pair spans could perform better for other classes of long RNA sequences, such as ribosomal RNA and long non-coding RNA. Note that although long non-coding RNA may be largely unstructured, local structured domains, or regulatory target sites could be located on these molecules making structure prediction interesting; for example for determining the accessibility of miRNA target sites [39].

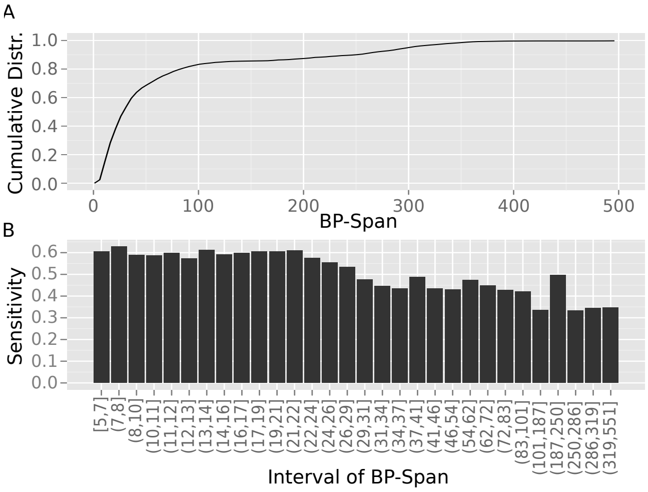


Figure 6.2. The distribution of base-pair spans and the quality of prediction with respect to span length. (A) The *bp-span* (x-axis) distribution for the CisReg dataset with the cumulative distribution given on the y-axis. (B) The sensitivity of base pairs (y-axis) for each base-pair span interval (x-axis). The intervals were distributed such that they contain roughly an equal number of base pairs. Figure taken from [P4].

6.5.3 Base-pair prediction accuracy decreased with span length

The choice of the locality parameter also depends on the prediction accuracy of base pairs with respect to their span lengths. For this evaluation, we used RNAfold as it allows all base-pair spans. The influence of the base-pair-span length on the sensitivity of the predictions is illustrated in Figure 6.2.B. We defined sensitivity as the fraction of all true base pairs within each *bp-span* interval that were predicted with probability $p(i, j) > 0.5$. Base pairs with a probability greater than 0.5 are called high-frequency base pairs and are contained in the centroid structure [37, 69, 152]. Base-pair prediction accuracy decreased with respect to span length; this was also published in [74, 85, 175]. The highest sensitivity of approx. 0.6 was achieved for *bp-span* < 30 nt, after which it dropped to around 0.45, and at *bp-span* ≤ 100 nt the sensitivity decreased further to around 0.35 (except an outlier at 0.5). The implications of this decrease are twofold: (1) The current nearest neighbour energy model [207, 315] is unsuited to the prediction of long-range base pairs and/or (2) the multi-loop energies are incorrect [64, 207, 208]. Our results indicated that an $L = 150$ represents a good balance between maximising the number of base pairs included in the predictions and minimising the inaccuracy of longer base-pair spans. A larger L did not increase the performance, probably due to the very few extra base pairs that could be predicted and the quality of those predictions becoming increasingly poor.

6.5.4 Structures are locally stable

The success of local folding approaches is based on the assumption that, in most cases, structures with short base-pair spans are locally stable and do not need the global influence of long-ranging base pairs to stabilise their formation. This condition is supported by the fact that small values for L performed only slightly worse than their more global counterparts (see Figures 6.1 and 6.8). In the search for *cis*-regulatory elements, maximum base-pair spans much smaller the real spans still predicted the local parts of the structure. The structural stability of local substructures was also stated in [74, 233]. These authors illustrated that in predicted, sub-optimal structures, most of the rearrangement occurs in the form of long-range connections, whereas the local substructures remain the same. Moreover, Higgs and colleagues have shown that, due to kinetics, short-range base pairs form more quickly [222]. Finally, the hierarchical evolution hypothesis, introduced in [25], could further support the initial formation of locally stable structures with short base-pair spans and the subsequent addition of longer-range connections.

6.6 Artificial window borders can be detrimental to structure prediction

The window-based approach, `RNAplfold`, computes base-pairing probabilities by averaging over subsequences, windows, of length W. On the one hand, averaging over independent windows reduces dependencies between two local structures with a distance greater than W; on the other hand, each window introduces two artificial RNA ends at the window borders. As the ends do not correspond to any real features of the RNA, these can lead to the following errors; with the appropriate care, these errors can be avoided.

6.6.1 Window borders were biased towards higher accessibilities

To investigate a possible bias introduced by folding independent (short) subsequences, we computed the average accessibility per position of the respective windows using `RNAplfold`.

Mean accessibilities for over 500, 000 sequence windows from 400 mRNAs, selected randomly from four species, are depicted in Figure 6.3. Nucleotides at the window borders showed considerably higher accessibilities than nucleotides near the window centres. This effect is preserved for the full range of observed GC-contents (Figure D.15) and is not particular to mRNAs (Figure D.16). Our expectation that most of the bias originated from external regions not enclosed by any base pair, as opposed to internal loops, was confirmed (data not shown).

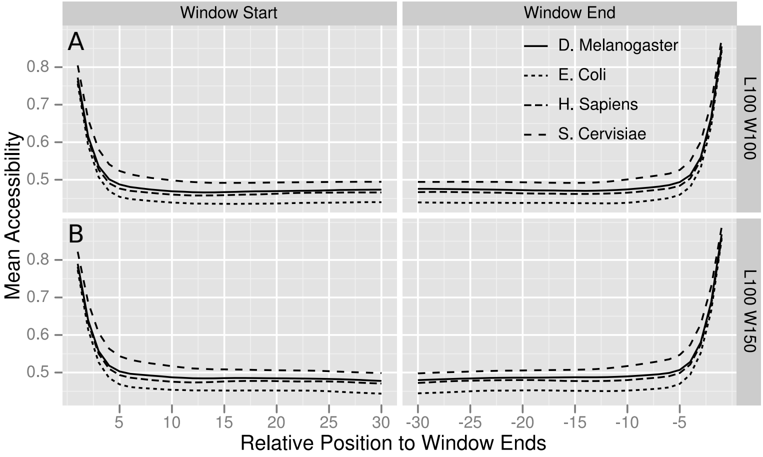


Figure 6.3. High accessibilities at window borders. Average accessibilities were computed per window position for 400 randomly chosen mRNAs from four species. Computations were done with `RNAplfold`, $L = 100$ and (A) $W = 100$ and (B) $W = 150$. Positions beyond approx. 10 nt at the window borders have equivalent average accessibilities. Figure taken from [P4].

6.6.2 The use of windows can also lead to biased base-pair predictions at window borders

The accessibility bias towards window borders affected the probabilities of base pairs when at least one of the two nucleotides (involved in a base pair) is situated within the border region. Consequently, long-range base pairs with both nucleotides within the outer regions were affected most (Figure 6.4.A). Two issues arise from window-based folding. First, the number of windows in the calculation of a base-pair probability is dependent on its span, i.e., probabilities of a base pair with *bp-span* $= l$ occur in $W - l + 1$ windows. Hence, the number of windows being averaged decreases linearly with increasing *bp-span*. Second, strong secondary structures tend to form in the central part of a window, leaving the remaining unpaired bases at the window borders available to pair with each other; crossing base pairs with internal unpaired bases are not allowed in secondary structure prediction, so the ends pair up (if possible), because each additional base pair minimises the overall free energy. In combination, when L is close to W, long-range base pairs within the borders resulted in skewed pairing probabilities, as they were not compensated by averaging over many windows.

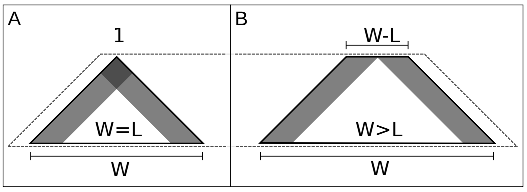


Figure 6.4. Illustration of folding windows. Regions affected by the border effect are shaded. (A) Same window size and maximum span. Long-range base pairs can be affected by both window borders. The base pair of maximal span is part of exactly one window. (B) Window is larger than the maximum span. Base pairs can only be influenced by one window end. Base pairs of maximal span can be part of multiple windows. Figure taken from [P4].

6.6.3 Border effects can be reduced by the appropriate choice of window size

The negative effect of having only few windows representing long-range base pairs was mitigated by setting a suitable window size W with respect to the maximum base-pair span L. When $W \geq L$, base-pair probabilities are averaged for at least $W - L + 1$ windows (Figure 6.4.B). In Figure 6.5, the dot plots from `RNAplfold` of a *cis*-regulatory element exemplify the border effect on long-range base pairs. For visualisation purposes, the sequences were folded with $L = 70$. For $W = L$, many base pairs with spans near L were assigned high probabilities while located in very short stems (Figure 6.5.A). For $W = L + 50$, most of the long-range base pairs either disappeared or were assigned much smaller probabilities (Figure 6.5.B). The base-pair probabilities for the target structure were not influenced by the parameter settings, due to their shorter base-pair spans. In our evaluations of different window sizes on both the `CisReg` and the `YeastUnpaired` datasets, W had little effect on the prediction performance as long as it was sufficiently larger than L. The current default parameter setting of `RNAplfold` is $W = L = 70$. In general, the default settings of computational tools are frequently used and in the case of `RNAplfold` the default, $W = L$, was applied in e.g. [170]. Note that on the other extreme, window sizes much larger than L diminish the positive effects of the window-based approach, namely to avoid dependencies between distant local structures. When W is equal to the sequence length, the window-based approach is the same as the approach for `Rfold` and `Raccess`. Varying the window sizes from $L+50$ to $3L$ did not influence the results significantly, however, the best results for `RNAplfold` were achieved using $W = L + 50$ (Figures D.17 and D.18). For all further evaluations we set the window size to $W = L + 50$, which allowed each base pair to be present in at least 51 windows (when the RNA length exceeds the window size).

6.6.4 LocalFold diminished border effects

While an appropriate choice of the window size mitigated some of the adverse effects of windowed approaches, the borders still affected the accessibilities up to the ten outer nucleotides of each folding window (Figure 6.3.B). Therefore, we developed `LocalFold` that reduced these border effects and we quantified the improvement of predictions performed on our datasets.

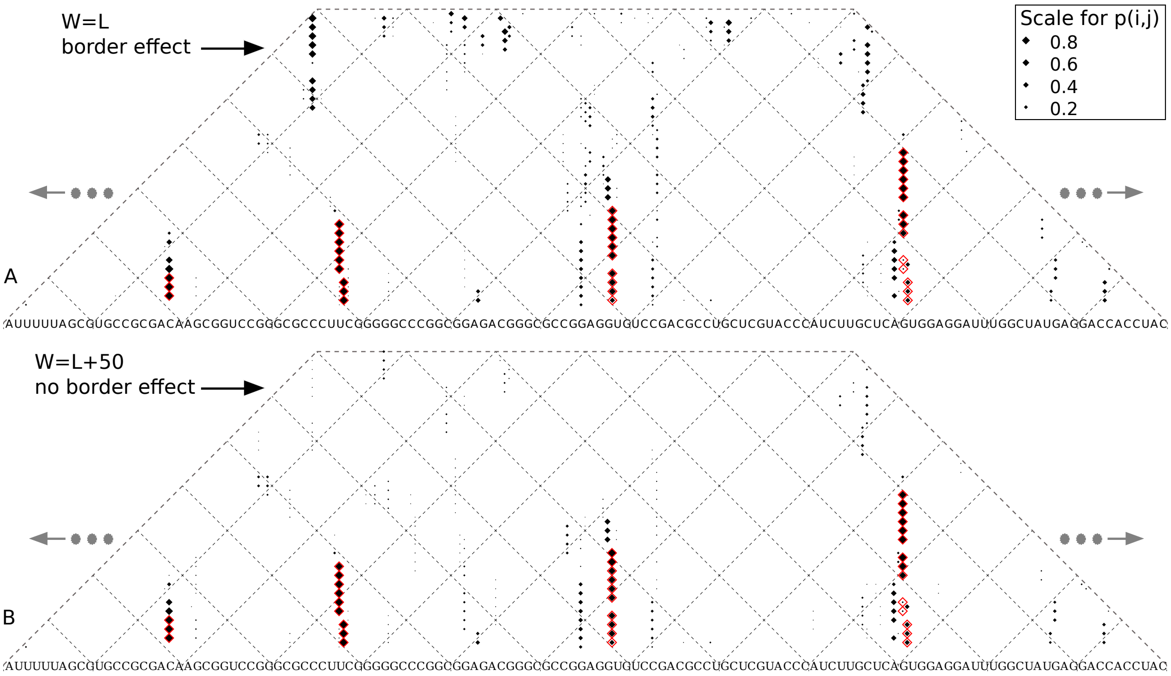


Figure 6.5. Probability bias for long-ranged base pairs close to the window size and their reduced effect. We see the original dot plots of the base-pairing matrices cropped for visual inspection to the nucleotide positions 5180 to 5291 of RF00435-U55047-1 in the CisReg dataset, which is a heat shock gene expression (ROSE) element. Base pairs of the target structure are marked in red. The size of each dot is relative to the probability of the base pair it represents and the nucleotides can be read by following the diagonal lines to the left and right. The incorrect long-range base pairs are much more likely when (A) $W = L$ instead of (B) $W = L + 50$. Figure taken from [P4].

In short, the biased regions at the window borders were not considered for the computation of accessibilities or base-pair probabilities. As the border effect was mostly independent of window size and maximum base-pair span (not shown), in LocalFold the first and last ten nucleotides in each artificial window (excluding real ends of the input sequence) were removed from the calculations. Note that LocalFold only removes the bias outliers from the window-average calculations and still produces probabilities for all positions of the nucleotide sequence (any length).

6.7 Performance comparison of methods

We compared the performance of the following secondary structure prediction methods applied to mRNA sequences: RNAfold (global), Rfold (restricted *bp-span*, base-pair probabilities), Raccess (restricted *bp-span*, accessibilities), RNAplfold (window-based), and our method LocalFold (reduced border effects). We investigated their performance on the CisReg and the YeastUnpaired datasets, hence, we quantified their predictions of both paired and unpaired bases, respectively. For the local folding methods, we applied the best parameter combinations (for each dataset) according to the previous analyses.

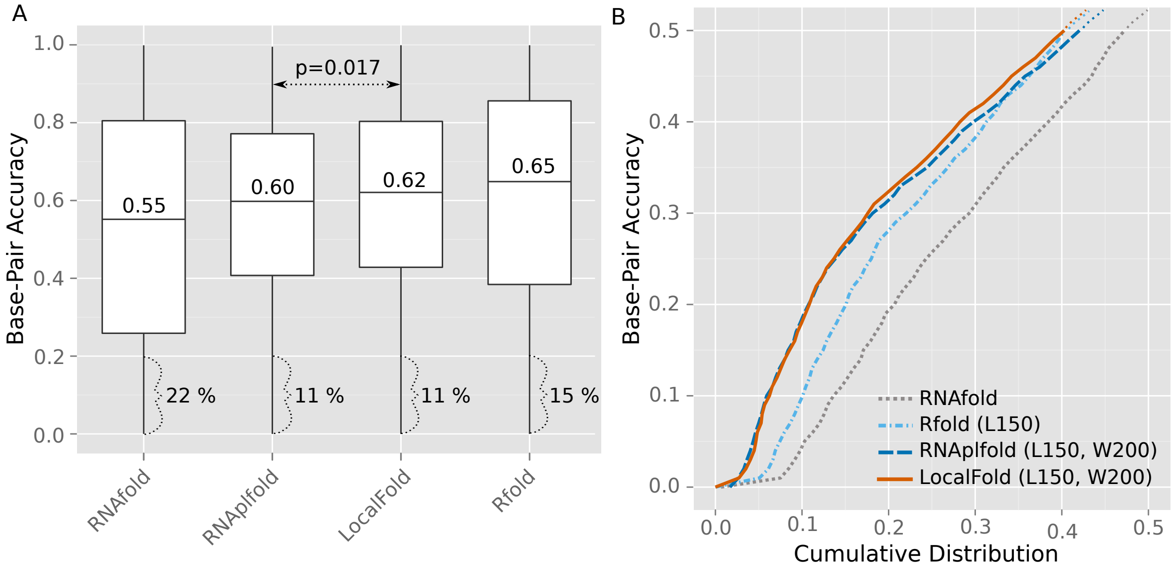


Figure 6.6. Comparison of structure prediction methods for the identification of *cis*-regulatory elements. Computations were performed with L=150 and W=200 (when applicable) on the subset of the CisReg data that have a max. base-pair span of 150 nt, including 2158 elements assigned to 90 Rfam families. (A) Comparison of the achieved accuracies as boxplots. (B) Cumulative distributions of the *bp-accuracy* up to 0.5 (y-axis) to highlight the prediction sensitivity. Base pairs with probabilities above 0.5 are contained in the centroid structure [37, 69, 152] and thus a *bp-accuracy* above this threshold implies a well defined target structure. The p-value was calculated with a two-sample Wilcoxon Rank Sum test. Figure taken from [P4].

6.7.1 Predicting *cis*-regulatory structures in mRNA

We compared the accuracies each method achieved for the base pairs in the CisReg dataset. For folding, we used sequences with up to 500 nt of context on either side of the elements (e.g. see Figure 6.7.B. Although many mRNA sequences are longer than 1000 nt, we chose this length because resource demands of RNAfold were too high for longer sequences. For the local folding methods, we applied the optimal values determined previously: maximum base-pair span $L = 150$ and window-size $W = 200$. To fairly compare RNAfold to the local folding methods, we used a subset of the CisReg dataset in which the elements had a maximum *bp-span* of 150 nt. This subset included most elements (2158 out of 2500) across 90 different Rfam families. This meant L did not exclude base pairs in the dataset from being predicted. In Figure 6.6, we summarised the *bp-accuracies* (Equation 6.4) resulting from each method. When comparing the median *bp-accuracy* in Figure 6.6.A, it increased from 0.55 (RNAfold), through 0.6 (RNAplfold), 0.62 (LocalFold), to a maximum of 0.65 (Rfold). These accuracies indicate that the target structures were clearly predicted as illustrated in Figure 6.5 in which the *cis*-regulatory element achieved a *bp-accuracy* of 0.65. Although Rfold achieved the highest median *bp-accuracy*, the method—together with RNAfold—exhibited a much greater variation in results than the window-based approaches, RNAplfold and LocalFold. While the boxplot indicated similar distributions for the latter two approaches, the accuracies for LocalFold were significantly higher than for RNAplfold ($p = 0.017$, two-sided, two-sample Wilcoxon Rank Sum Test). Both window-based approaches produced the most robust predictions; LocalFold and RNAplfold made fewer predictions in the lower *bp-accuracy* range, i.e., they were more sensitive (Figure 6.6.B). We considered a *bp-accuracy* ≤ 0.2 to mean the structure was not predicted: Rfold and RNAfold failed to predict 15 % and 22 %, respectively, whereas both RNAplfold and LocalFold failed in only 11 % of all instances. To show that these results were not biased by redundancies in the dataset, we evaluated the median accuracy per Rfam family (Figure D.19). Albeit some exceptions, the above trends remain the same for the individual families. Only for two families with large base-pair spans of 338 and 551 nt did global folding show a substantial improvement over the local folding methods.

6.7.2 Rfold has a decreased prediction performance at sequence ends

In the investigation of different context lengths for the local folding methods, Rfold exhibited a decreased performance for smaller contexts (Figure 6.7); the context length was defined by the number of nucleotides to either side of the regulatory element (Figure 6.7.B). Although the median *bp-accuracy* for Rfold was higher for the contexts of 200 and 500 nt, it performed worst for 100 nt. This, in combination with the greater variance for all Rfold predictions (evident from the quantiles in Figure 6.6.A), indicated that the prediction of correct structures at sequence ends is poor. A similar trend was observed in [171], where the authors reported decreased prediction for the ends of sequences up to four times the maximum base-pair span, i.e., a context of 600 nt for $L = 150$. Most *cis*-regulatory elements are situated within the UTRs of mRNAs and thus are frequently located at the sequence ends. Hence, poor prediction performance at sequence ends is detrimental for the prediction of *cis*-regulatory elements.

6.7.3 Evaluation of accessibilities in yeast data

In the previous analysis, we inspected the accuracy at which each method predicted a given secondary structure. The extent of incorrectly predicted base pairs was not explored. Here, we compared the performance of all methods on their ability to predict the accessibility of individual bases. As the accessibility of a base is defined as its probability of being unpaired, the probabilities of all possible base pairs involving this nucleotide are taken into account. Thus, incorrectly predicted base pairs can have a detrimental effect on this measure. We first computed accessibilities for each folding method. For the local-folding methods, we applied maximum base-pair spans (L) between 25 and 200 nt: the window size $W = L+50$ was used for the two window-based approaches. The quality of predictions for the YeastUnpaired dataset was evaluated by computing AUROC values for discriminating high- and low-rated nucleotides according to the PARS score [165]; these nucleotides achieved the clearest evidence for being paired or unpaired, respectively. Figure 6.8.A shows the results for the highest-ranking 1 % and the lowest-ranking 1 % nucleotides, comprising a set of approx. $80,000$ measurements. In most cases, an AUROC greater than 0.8 was achieved. Folding globally with RNAfold resulted in the third lowest performance, only the predictions of Raccess and RNAplfold using span $L = 25$ performed worse. LocalFold outperformed the other methods for all Ls.

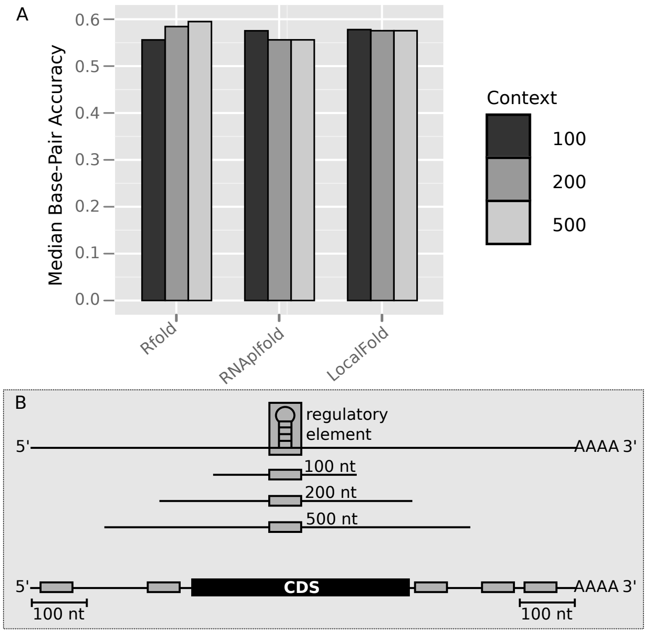


Figure 6.7. `Rfold` has increased problems predicting correct structures at sequence ends. `Rfold` is more sensitive to the context length and thus has increased problems predicting correct structures at sequence ends, also reported in [171]. (A) A comparison of the median *bp-accuracy* (y-axis) achieved by the local folding methods on sequences where the regulatory element is situated within contexts 100, 200 and 500 nt (`CisReg` dataset). (B) When the regulatory element is located at the sequence ends, a context larger than 100 nt is often unavailable. Thus, methods performing poorly for shorter contexts are not appropriate to identify those elements. Figure taken from [P4].

Even the worst result for `LocalFold` at $L = 25$ was significantly higher than for `RNAfold` ($p = 8.055 \cdot 10^{-8}$, Wilcoxon Rank Sum test using AUROCs derived from 1,000 bootstrap samples). The best prediction result was attained by `LocalFold` using $L = 100$ with an AUROC of 0.85. Larger L values resulted in comparable AUROCs, hence, the prediction of accessibility was stable for different parameter settings. The fact that `Raccess` was clearly outperformed by the window-based approaches on the `YeastUnpaired` data provides further evidence that the greater variance in its base-pair prediction performance (Figure 6.6) is detrimental.

6.7.4 Relative prediction performance was not influenced by transcript length

Finally, we investigated the influence of transcript lengths on the performance of the algorithms. For the analysis shown in Figure 6.8.B, we split the data into sequence length intervals and the AUROC for $L = 100$ was computed for each interval separately. The intervals were chosen to include roughly an equal number of sequences. We used the highest-ranking 10 % and the lowest-ranking 10 % of nucleotides so that each interval contained a sufficient number of sequences. While predictive performance fluctuated slightly for the intervals, we observed the same ranking of methods as seen in the previous analysis: global folding scored worst, the window-based approaches best. `LocalFold` scored marginally better than `RNAplfold` for

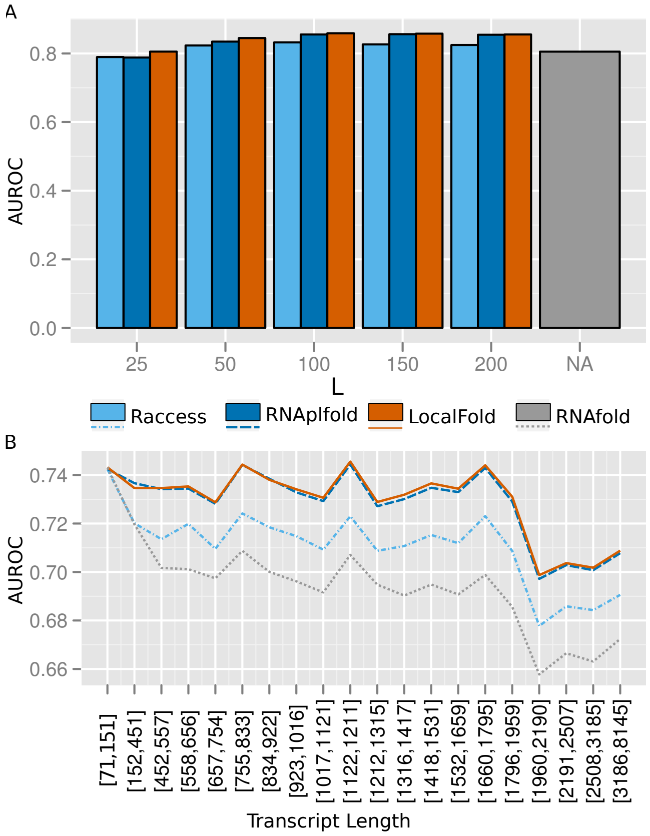


Figure 6.8. Comparison of AUROC values for separating high- and low scoring nucleotides of the YeastUnpaired dataset. (A) Effect of the parameter L was evaluated for $W = L + 50$ including only the 1 % highest and 1 % lowest scoring nucleotides. (B) Using the best parameter combination ($L = 100, W = 150$), we show the dependency of the transcript length on the prediction quality. Here the 10 % highest-ranking and 10 % lowest-ranking nucleotides were included. Each interval contains roughly the same number of sequences. Figure taken from [P4].

most intervals and both consistently outperformed Raccess. Overall, performance dropped slightly for sequences longer than 2,000 nt. The fluctuations in performance were mirrored by all methods, probably due to the quality or properties of the underlying data.

6.7.5 CisReg and YeastUnpaired data showed similar results

We observed similar results for both of the analysed datasets. The YeastUnpaired dataset was generated in *in-vitro* conditions, whereas, the structured *cis*-regulatory elements in the CisReg dataset consists of experimentally verified regulatory structures with post-transcriptional functions *in vivo*. The fact that the results are comparable between two independent datasets supports their overall quality and highlights their validity and generality.

6.8 Conclusion

To benchmark the performance of mRNA secondary structure prediction, we generated a large curated set of *cis*-regulatory elements and introduced *bp-accuracy* to measure how accurately a local structure was predicted. Furthermore, we evaluated accessibility predictions using transcript-wide structure-probing data. Prediction accuracy was affected by the following algorithmic assumptions and parameter combinations:

1. The optimal base-pair span parameters were dataset dependent, but similar, at $L = 150$ for the `CisReg` dataset and $L = 100$ for the `YeastUnpaired` dataset. Within a range of 100–150, differences in performance were minimal. This range reflects the distribution of base-pair spans for known structures.

2. The use of sliding windows allows for more locality than the mere restriction of base-pairs spans. Windows, however, introduced a prediction bias at each artificial border. Windows with $W = L$ caused unusually high base-pairing probabilities of long-range base pairs. This was was resolved by setting $W = L + 50$.

3. Setting the larger window size ($W = L+50$) did not remove the bias of high accessibilities (single-strandedness) at the window borders. Therefore, `LocalFold` was developed to diminish this bias which resulted in a consistent improvement compared to the other methods.

The greater improvement in results was observed for the `CisReg` data (base pairs) in comparison with the `YeastUnpaired` data (single-strandedness).

In addition to having much faster runtimes, we present clear quantitative and qualitative evidence that local folding methods outperformed the global approach. The advantage of local folding is that the majority of base pairs have short base-pair spans and that local structure can be predicted without the stabilising effects of long-range connections. Moreover, the reduced accuracy in the prediction of long-range base pairs meant that local folding was better than global folding at determining secondary structure in long RNAs.

CHAPTER 7

CRISPR structure prediction: the influence of context on the formation of local structure motifs

An RNA is frequently modified, transported, or processed by a mechanism that involves the binding of a *trans* factor to a local structure motif located in the longer RNA sequence. Examples of such local structured motifs are the *cis*-regulatory elements within mRNA transcripts in the `CisReg` dataset that were used in Chapter 6 to benchmark the performance of local-folding algorithms (Section 6.3). A significant part of this thesis is dedicated to the characterisation of other such local RNA structure motifs that are found in many CRISPR-Cas immune systems in prokaryotes (published in [P3, P5, P7, P10]).

In Part II, we used the conservation of CRISPRs to determine structure motifs and sequence families that characterise properties of the repeat, which might influence the binding affinity of associated Cas proteins. When the full CRISPR array is being processed into its many mature crRNA species, each repeat instance is, however, imbedded within varying sequence contexts (spacers). The functional structure motif within each repeat instance may thus be stabilised or destabilised by surrounding structure formations. In this chapter, we calculate structure stability profiles of each repeat instance in an array to measure the influence of the surrounding sequence context. For the first time, we provide biological evidence that the sequence context surrounding a repeat can indeed reduce structure-motif stability and consequently inhibit crRNA processing by forming stable base pairs with the repeat that are in conflict with the functional structure motif—despite that the repeat sequences are identical. We assume that, in general, spacers for native CRISPR arrays are selected such that they do not disrupt the formation of the functional structure motif in the repeat. Following this assumption, we present a method to predict stable structure motifs of repeats across single CRISPR arrays; we show that the most stable structure (on average) in the array may deviate from the MFE structure of just the repeat sequence.

A CRISPR array can contain from 3 to over 1,000 repeat instances (see Table 3.1, Section 3.1) and can comprise many kilobases—similar to mRNA transcripts. Thus, global structure-prediction approaches might not provide accurate results when applied to the entire array. We used the knowledge gained in Chapter 6 to set appropriate parameter values.

7.1 Computation of structure-stability profiles in CRISPR arrays

Once the characteristic stem-loop motif of a specific CRISPR has been determined, its stability at every repeat instance in the CRISPR array can be measured by calculating its structure accuracy, *bp-accuracy* (Equation 6.4, Section 6.2). Put simply, the *bp-accuracy* is the average base-pair probability of all base pairs in a given structure. Stable structures, i.e., structures that form with a high probability, have a high structure accuracy.

Structure-stability profiles (Definition 7.1) were computed by first performing a local structure prediction on the entire array using `RNAplfold` [16] with the window-size and base-pair–span parameters set as $W = 150$ and $L = 100$, respectively; no lonely base pairs (option `--noLP`) were allowed. Base-pair probabilities are stored in a local dotplot (see Section 2.5.5).

Definition 7.1. *Let $\mathcal{M} = (N, B)$ be the given repeat structure and $\mathcal{L} = (1, \ldots, n)$ denote the starting positions of each repeat instance in the array (with n repeats in the array). Then $\mathcal{M}^l = (N^l, B^l)$ is the repeat structure at position $l \in \mathcal{L}$ and the respective set of base pairs is given by $B^l = (i + l - 1, j + l - 1), \forall (i, j) \in B$ (i and j are 1-based indices). The* ***structure-stability profile*** *is given by $\mathcal{P} = (\textit{bp-accuracy}(\mathcal{M}^1), \ldots, \textit{bp-accuracy}(\mathcal{M}^n))$. The bp-accuracy($\mathcal{M}^l$) can simply be calculated by looking up the base-pair probabilities of B^l in the dotplot and computing their average*[1]*.*

As an example of CRISPR-structure stability, we looked into the two CRISPRs in *Methanosarcina mazei* Gö1 [P7]. Figure 7.1 depicts the MFE structure of the consensus repeat sequence in part A and repeat-sequence variants at CRISPR loci 1 and 2 in part B. The repeats at both loci are generally identical, however, some repeat instances contain point mutations (highlighted in pink). We note that the MFE structure is conserved across the array, i.e., none of the mutations disrupt the MFE structure. Furthermore, this structure was verified using *in-vitro* structural probing in [P7]. The structure-stability profile in Figure 7.1.C displays a huge variance in the MFE structure stability across the array at CRISPR locus 1; especially at positions 16 and 24, the MFE structure is sequestered by the neighbouring spacer sequences[2].

[1] Note that since local folding was performed, the probabilities are, in fact, average base-pair probabilities across multiple sliding windows; their square root is saved in the dotplot for improved visualisation.

[2] The same variability is observed for CRISPR locus 2, but was not depicted.

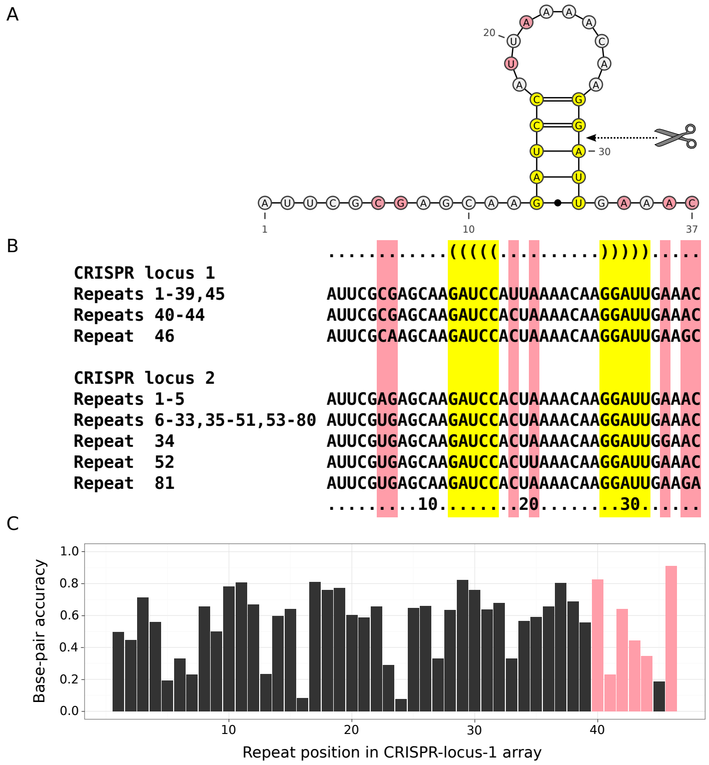


Figure 7.1. CRISPR structures in *Methanosarcina mazei* Gö1 and the variability in structure stability. (A) Shows the minimum-free-energy structure of the consensus sequence of all repeat instances in both CRISPR loci 1 and 2. The structure was verified using *in-vitro* structural probing in [P7]. Pink nucleotides are positions of point mutations in the repeats of both CRISPR loci. (B) All repeat variants are shown for both CRISPR loci in *M. mazei.* Colums with point mutations are highlighted in pink and the bases involved in the base-pairing of the MFE structure are highlighted in yellow; the structure is conserved across all repeat instances. (C) The structure-stability profile of locus 1 shows that a large range in structure accuracies exist: some repeat positions form stable structures, whereas at other positions the MFE structure is very unlikely to fold due to influences of the surrounding sequence context. Figure modified from [P7].

7.2 The efficiency of cleavage at repeats in a CRISPR array

In Chapter 5, we established a large variance in the stable population of crRNAs processed from a single array. There are two factors that could explain these differences in expression: (1) the processed crRNA is not protected by associated Cas proteins and thus quickly degraded, and (2) the recognition and cleavage of the repeat is not either efficient or inhibited. Using *Synechocistis* sp. PCC6803 as an example organism, we provided evidence that an ill-chosen spacer could possibly lead to faster crRNA degradation (c.f. Section 5.5); now, we provide experimental proof that the sequence context, surrounding a repeat, can sequester the formation of the repeat structure motif and that this inhibits repeat cleavage.

7.2.1 Experimental analysis of repeat cleavage events

The genome of *Synechocistis* sp. PCC6803 contains three proteins that are homologous to the Cas6 family of endoribonucleases known to cleave CRISPRs during crRNA maturation: Cas6-1 is in the vicinity of CRISPR1, and Cas6-2a and Cas6-2b are close to CRISPR2

(Figure D.13). *In-vivo*, knock-out experiments showed that Cas6-1 processed the CRISPR1 array and that Cas6-2a was involved in processing mature crRNAs from the CRISPR2 array[1] [P10]. To investigate single, repeat-cleavage events, we required a soluble protein for *in-vitro* experiments. No purification of the Cas6-2a protein was possible—as it was not soluble—but purification worked for Cas6-1. Interestingly, Cas6-1 could process both CRISPR1 and CRISPR2 arrays *in vitro* [P8]; an observation that was not clear from the *in-vivo* experiments [P10]. In the subsequent analysis, the ability of Cas6-1 to cleave repeat instances in the CRISPR2 array was determined.

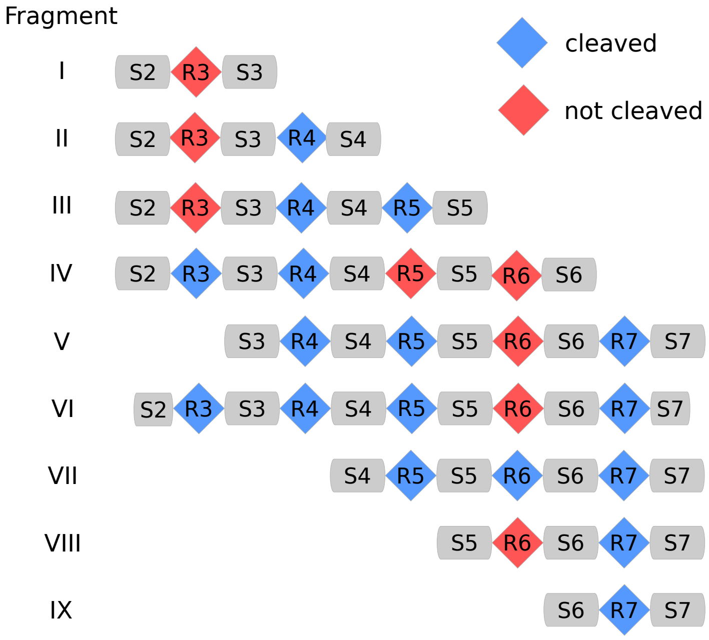


Figure 7.2. Repeat cleavage in different subsequence fragments of the CRISPR2 array in *Synechocistis*. In each of the nine experiments I–IX, the represented subsequence fragment of the CRISPR2 array was incubated with the purified Cas6-1 protein. The presence or absence of all possible cleavage fragments were determined by a subsequent northern blot analysis. If an observed length could be allocated to a cleavage product where either or both ends resulted in a cleavage within the repeat, then this repeat (or the repeats corresponding to both ends) were cleaved (blue). If such a length was not observed, then no cleavage occured (red). Figure modified from [P8].

Processing of repeats in the CRISPR2 locus was measured using a simple northern blotting approach: the purified Cas6 protein was added to the CRISPR2 array sequence, *in-vitro*, and the lengths of accumulated processing products were determined in northern blots. Albeit being cheap and fast, the caveat of this approach is that when considering the entire array, it would be difficult to map most lengths to a unique product. Therefore, smaller fragments of the CRISPR2 array were analysed. In addition, by selecting sub-sequences of the array, repeat cleavage could be investigated in various sequence contexts. The cleavage of repeats R3–R7 from CRISPR2 was investigated using nine different subsequences with a varying range of sequence contexts (see Section D.3.2 for the exact sequences). For each fragment, a cleavage experiment and subsequent northern blot was performed: each experiment is

[1] No function was found for Cas6-2b.

denoted by a Roman numeral from I–IX[1]. Each fragment was decomposed into its theoretical processed products, assuming all combinations of cleavage events in the repeat. For each repeat in a fragment, the presence or absence of the theoretical processing product that begins or ends at a cleavage site in that repeat, was observed: if a product was detected, cleavage of that repeat occured, if not, the repeat was not cleaved. The fragments and the events of repeat-cleavage per fragment is illustrated in Figure 7.2.

7.2.2 Calculation of repeat-structure stabilities

For each sequence fragment depicted in Figure 7.2, we calculated structure profiles using the repeat structure from Figure 7.3.A as published in [P10]. The structure accuracy, i.e., the measured stability of the given structure, for each repeat instance was taken from the structure profiles, which were calculated as described in Section 7.1.

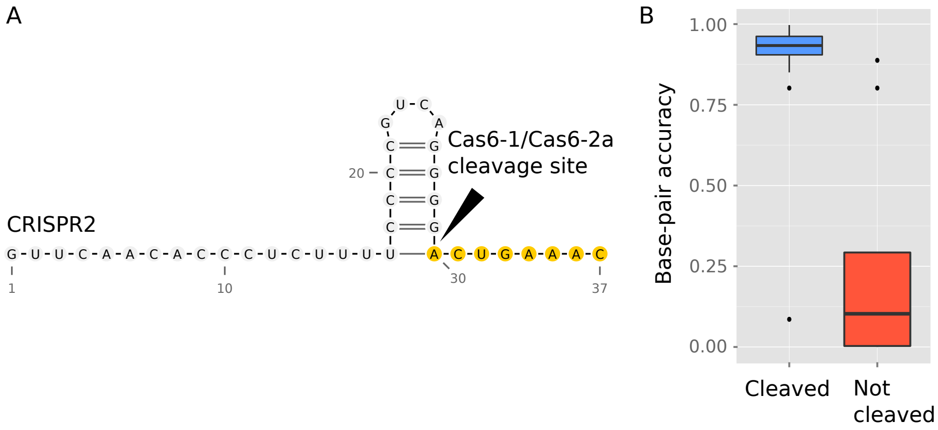


Figure 7.3. Structure motifs with a low measured stability were not cleaved at the CRISPR2 locus in *Synechocistis* sp. PCC6803. (A) The repeat structure motif, as published in [P10]. The structure belongs to structure class M5 in CRISPRmap (Chapter 3). The black wedge indicates the cleavage site of Cas6-1 (or Cas6-2a) and the yellow nucleotides mark the 8-nt tag remaining on the mature crRNAs. (B) Cleaved repeat instances (blue) form stable motifs structures and a high base-pair accuracy, *bp-accuracy*, is measured (x axis); in contrast, repeats that are not cleaved (red) do not form a stable motif structures as can seen by the low base-pair accuracies. Figure modified from [P8].

7.2.3 Repeats with a low structure stability are not processed

There was a clear-cut difference in the calculated stabilities of the structure motif of CRISPR2 that is recognised by Cas6-1 between successful and unsuccessful repeat cleavage (Figure 7.3): repeat instances that were cleaved had very high base-pair accuracies (i.e. measured stabilities), whereas, instances that were not cleaved displayed a marked decrease in *bp-accuracy*. This trend was also shown in dotplots, where the average base-pair probability for cleaved vs. uncleaved repeat instances was compared. We also included the average base-pair probabilities of the surrounding spacer sequences (that have variable sequences) using the mode spacer length in CRISPR2 (see additional Figure D.20). The dotplots show

[1] The exerimental details are not part of this thesis and are given in [P8].

that repeat loci, which are not cleaved, form stable structures with their surrounding spacers that interfere with the functional structure motif; thus, the motif appears with reduced base-pair probabilities than in cleaved repeats.

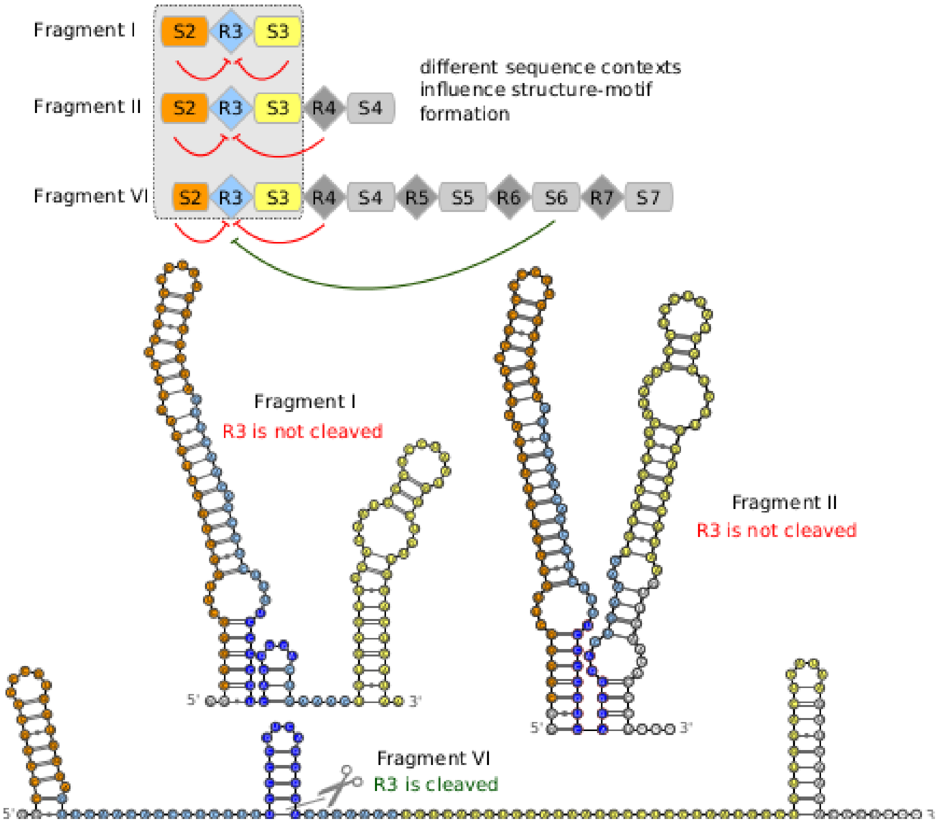


Figure 7.4. Example of how different sequence contexts can influence repeat folding and affect crRNA processing. In the three fragments I, II, and VI, the cleavage of R3 was analysed. In all fragments, R3 is surrounded by the same spacers, except that in fragment VI only half the spacer is present, but the fragment is extended further by a different number of repeat-spacer units. Each fragment is a real subsequence of the native CRISPR2 array. From a local structure prediction performed on each fragment separately, we extracted the base pairs with an average probability greater than 0.5 and assembled these into the three most-likely structures for the region S2-R3-S4. We see that the respective context sequences have a clear impact on the R3 structure: both fragments I and II form stable base pairs with the surrounding spacers and are not cleaved; the surrounding context in fragment VI is largely unstructured such that the functional repeat-motif structure is stable and in this case, cleavage occurs. Figure taken from [P8].

Notably, the pattern of repeat cleavage is not black and white: some repeat instances exist that are sometimes cleaved and other times not—even though the directly neighbouring spacer sequences are the same. This implies that long-range influences occur that either stabilise the repeat structure or favor alternative structures that are in conflict with the functional binding-motif structure. An example is given by repeat R3 in Figure 7.4: Despite identical neighbouring sequence context, repeat R3 is not cleaved in fragments I and II but is cleaved in fragment VI. Both I and II fragments are not cleaved in R3, probably because long, stem-loop structures, forming base pairs with spacers S2 and S3, prevent the structure binding motif from forming. In fragment VI, the first half of S2 is missing, therefore, only a smaller stem-loop is formed here, allowing the structure motif to form. In addition, structures further downstream make S3 unstructured and may allow the Cas6 protein better access for binding and cleavage.

7.3 CRISPR-specific, context-based structure prediction of repeats

In Part II, we used evolutionary conservation to detect CRISPR structure motifs. The entire scope of evolutionary diversity of bacteria and archaea has not yet been captured, therefore, in individual cases, there is no convervation information available. In such instances, an alternative structure-prediction approach is required. The fact that repeat instances with unstable structure motifs are not always cleaved means that an efficient CRISPR locus would be one where the functional structure motif is the most dominant formed structure. We exploit this assumption in the following repeat-structure prediction approach.

The general practise in the search for the functional CRISPR repeat structure is to compute the MFE structure of a single repeat sequence. The repeat is not transcribed as a single unit, however, but is located on a transcript in the context of other spacers and repeats. These flanking sequences can impact the structure formation such that sub-optimal repeat structures could be preferred over the MFE structure. Although the MFE prediction is frequently correct due to highly stable stem-loop structures with many GC base pairs [P3], we show that this procedure may not always be accurate. Our structure-prediction approach, tailored specifically to CRISPR features, includes the entire array sequence and determines the most stable structure formation within that context (illustrated in Figure 7.5). The following steps resulted in the more accurate repeat-structure prediction.

1. The most probable repeat-structure candidates can be determined by visually inspecting the base-pair probability matrix (i.e. dotplot in Figure 7.5.B) of the repeat sequence as calulated by `RNAfold` [138]; the alternative is to calculate suboptimal structure candidates using `RNAsubopt` [344][1]. Usually `RNAsubopt` produces very many suboptimal structures (w.r.t. the Gibbs free energy), however, since CRISPR repeats are so short, the numbers are limited.

2. To determine the influence of the context sequence on each repeat sequence location, we predicted the structure of the entire CRISPR array. Due to the length of long CRISPR array sequences and unknown contexts that could arise through intermediate processing steps, we used the local folding approach `RNAplfold`[2] [16]. The locality parameter settings for the window size (W) and the maximum base-pair span (L) were taken from Chapter 6.1, published in [P4].

3. Subsequently, the submatrices for each repeat instance were averaged to form an average dotplot for the repeat structure (see Figure 7.5.C). The dotplot visualises the average

[1] The older Vienna package version 1.8.4 was used, with parameters '`-p -d2 -noLP`' for `RNAfold` and -s -noLP for `RNAsubopt`. Omitting the option '`-p`' for `RNAfold` calculated the MFE structure.

[2] Vienna package version 1.8.4, options '`-noLP -W 150 -L 100`'.

base-pair probabilities for the repeat sequence for all occurrences in the array and includes the influence of the context.

4. The candidate from (1) with the highest structure accuracy in the average dotplot from step (3) represents the most probable structure for that CRISPR array. This is the structure that has the highest probability on average across each repeat position. Thus, it is likely to form more frequently at repeat locations than the other candidates. The chosen candidate with the highest accuracy can usually be easily identified in the average dotplot, due to its greater base-pair probabilities and therefore larger dot sizes (blue structure in Figure 7.5.A).

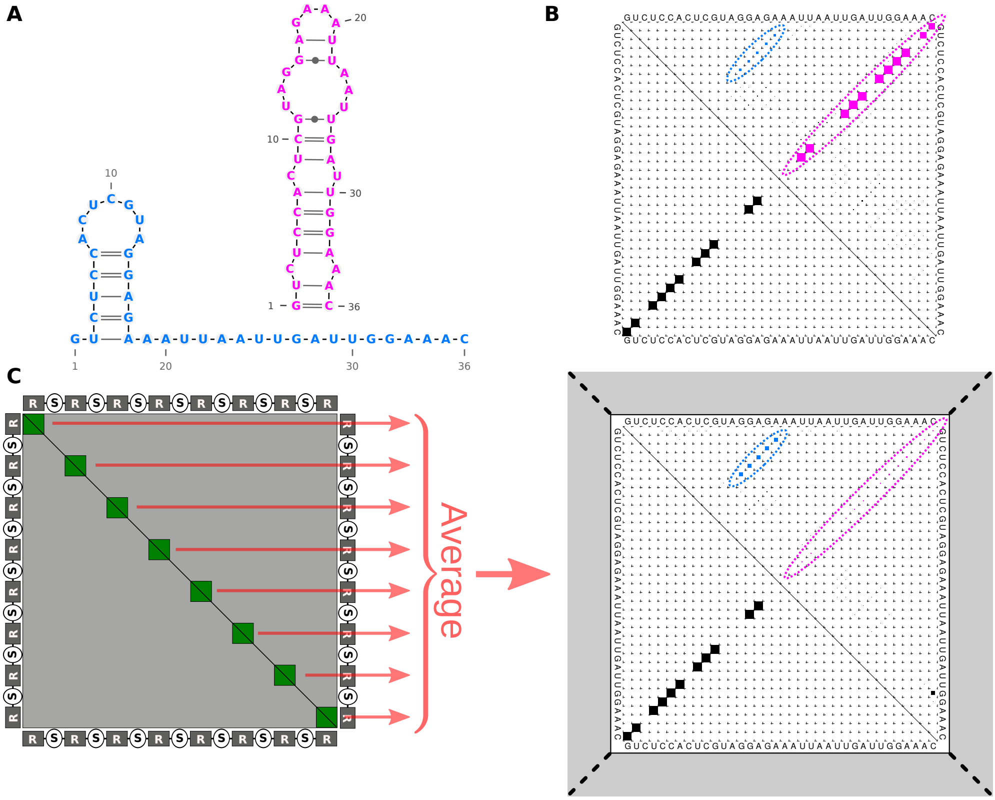

A
B
C
Average
GUCUCCACUCGUAGGAGAAAUUAAUUGAUUGGAAAC
G-U—A-A-A-U-U-A-A-U-U-G-A-U-U-G-G-A-A-A-C
1
10
20
30
36
R
S

Figure 7.5. Comparison of structures resulting from the commonly used MFE prediction to our CRISPR-specific context-based approach. Folding process is exemplified for CRISPR3 from *Synechocistis* sp. PCC6803 from [P10]. (A) The two most stable structure candidates; the MFE structure is in magenta. (B) The base-pair probability matrix, as computed by RNAfold [138], for the repeat sequence where the MFE structure is in the lower triangle and the two structures from (A) are clearly marked in the upper triangle. (C) Our approach: repeat structure in context. To analyse the influence of the context, we calculated the base-pair probability matrix for the complete array (R = repeat, S = spacer). The preferred structure in the context was determined by averaging the sub-matrices associated with the repeats. When the repeat was folded in its sequence context, the magenta structure nearly disappeared and the blue structure, which looks more like other known CRISPR structures, was more probable. Figure taken from [P10].

With this approach, we identified a repeat structure for CRISPR3 (blue structure in Figure 7.5.A) that resembles native CRISPR structures [P3] much more closely than the MFE structure (magenta structure in Figure 7.5.A). Whereas, for CRISPR1 and CRISPR2, the repeat MFE structure was also the most probable within its context [P10]. Future work in collaboration with experimental molecular biologists would be to verify that this structure indeed guides Cas binding.

7.4 Conclusion

Recently, CRISPR-Cas systems have been used as a basis for a new genome-editing technology [145, 203, 328]. Currently unstructured CRISPRs of type II are being used in an artificial setting using single guide RNAs. In addition, it is perceivable to use artificial CRISPR arrays in the future that are designed to target multiple locations simultaneously. The influence of sequence context on structure motifs becomes highly relevant when designing such artificial CRISPR arrays, especially when multiple repeat-spacer units are involved: it is important to assess the stability of the functional structure motif for each repeat instance in an artificial CRISPR array. The structure is one of many factors that influences cleavage efficiency. It is interesting to note that many crRNAs did not lead to a successful defence of the invader in *Haloferax volcanii* [P5, P11]. Thus, it is evident that artificial constructs that include multiple repeat-spacer units must first be optimised or screened first for cleavage efficiency.

The more general, computational conclusion of this chapter is that the sequence context surrounding a structured regulatory motif can significantly contribute to structure formation: although most repeat instances in a CRISPR array are identical, the stability of a repeat structure motif can vary between each locus, and even alternative structures can be formed (see Figures 7.1.C and 7.5). This is a very important observation for the application of structure prediction algorithms to structured, regulatory RNA. A naive way to predict the secondary structure of a local region of interest, embedded within a larger transcript, would be to extract this region and fold it globally. Results in this chapter demonstrate that this naive approach may not always provide the functional structure; however, it is a valid approach if conservation of base pairs is used for a more informed structure prediction—as was done for identifying the CRISPR structure motifs in Part II. In Chapter 6, we determined that (1) there are border effects at artificial sequence ends and (2) the range of the influence of the context on a local structure is generally about 100–150 nt. Thus, taking the entire transcript, or sufficient context (see the *viewpoint notion* in Part V, Section 9.2.1), and folding it with a local folding approach, such as `RNAplfold` can take influencing context sequence into account while ignoring the detrimental effect of predicting long-range base pairs.

Part V

Characterising regulatory recognition elements

Part V: Characterising regulatory recognition elements

There are plenty of acquaintances in the world; but very few real friends.—Chinese proverb

Regulatory non-coding RNAs (ncRNAs) and RNA-binding proteins (RBPs) are *trans* factors that bind to local, regulatory recognition elements (RREs), frequently found in mRNAs to regulate their expression (see Section 2.7). Interactions between RNA and other molecules in the cell occur all the time by chance. It is the affinity between the *trans* factor and the RRE, however, which determines the strength of the interaction. The stronger the interaction, the more likely it is that the interaction initiates a regulatory process.

We are particularly interested in determining interactions between miRNAs and their recognition sites (MREs) on target mRNAs, whose expression is generally down-regulated subsequent to miRNA binding (Section 2.2.1). One of the earliest approaches for a computational detection of miRNA targets is `MiRanda` that was published in 2004 and basically uses a strict seed filter for finding MRE sites. In the last decade a manifold of further prediction approaches have been developed. Their overall accuracy, measured e.g., in precision and recall[1], is low and only marginal improvements were achieved since the first attempts. For example, on transcriptome-wide data comprising measurements of protein-expression-level changes in response to miRNA transfection or miRNA knock-down [285], published methods have performed very poorly. The highest recall achieved was about 45 % by using only a simple seed search of complementary matches to the region 2–7 nt of the mature miRNA sequence at a very low precision [3, 250, 251]. At relatively high precisions of roughly 49 %, sensitivities of below 15 % were achieved and at even higher precisions of 60 %, sensitivities were below 0.05 % [3,250,251]. To summarise, the accuracy of published prediction approaches is not sufficient for a robust application for biologists searching for candidate MRE sites. Therefore, biologists have resorted to using variants of the `CLIP-seq` protocol to detect MRE sites (Section 2.7.1). Although these have delivered good results in general, experiments do

[1] Recall is also referred to as sensitivity.

not always show a high overlap in detected sites since they are constrained to measurements of interactions that occur at the time of measurement, and detected MREs are specific to the tissue being evaluated. Moreover, mapping difficulties increase the number of MRE sites that remain undetected. Thus, many binding sites are still missed by these experimental approaches; an accurate computational detection is imperative for complementing such experiments.

In this part, we took preliminary steps to improve computational prediction of MRE sites. First in Chapter 8, we performed an empirical analysis of the statistical importance of sequence and structure characteristics of regions flanking MRE sites to extend our current model of functional miRNA-MRE interactions. Second in Chapter 9, we developed a natural and highly flexible encoding RNA that is processed by an efficient graph kernel to generate high-order features for machine-learning approaches to model any class of RRE interaction. In particular, we focus here on its application to modelling miRNA-MRE interactions; a previous application to RBPs was shown to be very successful [P6].

CHAPTER 8

The significance of sequence and structure flanking miRNA-recognition elements

Most prediction approaches limit themselves to assessing features within the boundaries of MRE sites [213, 215, 305, 310, 342]. To improve the detection of MREs, some studies have searched beyond the direct MRE site to explore the flanking sequences. In particular, structural accessibility and nucleotide composition of the flanking sequences were explored [109, 143, 164, 171, 234, 303]. A commonly assumed model of binding is that the MRE site has to be accessible for miRNA binding and the direct context should be accessible to allow room for the larger Argonaute (AGO) protein, which is bound to the miRNA sequence [164, 323]. Therefore, the accessibility of regions of various lengths around the miRNA (and siRNA) binding sites have been assessed for their statistical significance [109, 143, 164, 171, 234, 303]. These regions mostly overlap with the MRE site and thus assume a single binding event. To explore the possibility of additional binding factors that may influence miRNA regulation, we measured the significance of sequence and accessibility signals in regions that do not overlap with MREs[1].

8.1 MicroRNA interaction data

To assess the significance of accessibility around the MRE sites of plants, we curated a high-quality dataset of miRNA interactions with mRNAs in the model plant organism *Arabidopsis*

[1] Please note that a similar analysis was performed later in 2011 in [171], but, to the extent to our knowledge, the work presented here was done prior to any publications on non-overlapping regions with MREs in 2009. Due to time constraints, we did not pursue publishing this work previously, however, it motivated many subsequent ideas in this dissertation to include more context when modelling or analysing RRE motifs.

thaliana based on experimental evidence. The set contains exact hybridisation patterns for 110 functional and 114 non-functional miRNA-MRE pairs[1].

The functional set was extracted from degradome data performed by German and colleagues [102]. In particular, deep sequencing was performed in two *A. thaliana* cell lines to identify cleavage products of miRNA-directed degradation of target mRNAs: wild type `col-0` and the mutant `xrn4-/-`. An overrepresented abundance of reads that start or end at the same position indicate a cleavage site. Such cleavage sites that lie within the reverse complement sequence of known mature miRNAs were considered as evidence for a target site[2]. German and colleagues provide the miRNA and the target mRNA accession numbers, reads from the `RNA-seq` experiments that show the cleavage by the miRNA-RISC complex and contain half of the MRE sequence, the abundances of these reads in the two cell lines, and the exact cutting position. The following steps were performed to filter and extend these data to provide more detail on the exact hybridisation pattern of each interaction, and subsequently, accurate boundaries of the MRE site:

1. Most miRNA genes are processed into multiple mature sequence variants; a new interaction entry was made for each variant that fit to the data.

2. To identify the target site on the mRNA, a `BLAST` [5] search was performed with the reverse complement of each miRNA sequence. The best hits that coincided with the given cutting points were used to identify the exact MRE position and its boundaries.

3. All miRNA targets were removed that did not contain signature MRE reads for *both* cell lines to maintain a high quality of the data.

4. A prediction was made of the hybridisation between miRNA and mRNA target site by `IntaRNA` [34]. Due to the fact that the miRNA and its MRE share a great degree of complementarity, these hybridisation predictions should be accurate.

It is a very difficult task to gather a set of predicted hybridisations between miRNAs and mRNAs that are not functional in the cell, i.e. non-functional MRE sites, and no such dataset exists that is large enough for a statistical analysis. Most non-functional sites found in the literature are due to mutation experiments and are therefore not native. We generated a set of non-functional interaction pairs that share similar hybridisation patterns to functional interaction pairs based on experimental evidence that the non-functional pairs do not degrade the target. First, the results from the `Target Search` prediction method, which is part of the Web MicroRNA Designer `WMD3` [236], was used to predict potential target sites and these were filtered according to two criteria. (1) All verified mRNA targets given for each miRNA from the `ASRP` [120] database were removed, and (2) the expression data given by the `ASRP` database was used to delete those mRNAs from the set that showed more than 5 % knock down in the dicer mutant `dcl1-7` in comparison with the wild type `col-0`. In addition, all

[1] This dataset was generated in October 2009 and all data were downloaded from public databases at this time.

[2] These data include only MRE sites that result in mRNA cleavage and not inhibition.

pairs were removed that showed no expression of either the miRNA or the mRNA. For the GEO (gene expression omnibus) accession numbers of the expression data, see Table D.21.

8.2 Signals of higher accessibility downstream of MRE sites

We compared accessibility signals between the curated functional and non-functional MRE sites using a sliding window approach in the surrounding sequence context. In this way, not only the MRE sites but regions independent of the miRNA binding site were assessed.

8.2.1 A sliding window approach to assess accessibilities around MREs

Because miRNAs and their MREs are of variable length, the miRNA seed sequence was used as an anchor to align the MRE flanking sequences. We define the nucleotide opposite the first nucleotide in the miRNA sequence as position zero. Flanking sequences 200 nt up- and downstream of position zero were extracted from the target mRNAs for both the functional and the non-functional set described in Section 8.1. A single sequence which includes either a functional or a non-functional MRE site is denoted as R. Accessibilities were computed for R using `RNAplfold` from the Vienna Package version 1.8.4 and the following settings for the locality parameters: a window size (W) of 100 nt and a maximum base-pair span (L) of 50 nt[1]. We set $U = 10$ as the `RNAplfold` parameter for the maximum length of accessible regions. With this setting, `RNAplfold` calculates for all $u \in \{1, \ldots, U\}$ the mean probability $\overline{pu}(i, j)$ with $j - i + 1 = u$ that the subsequence $R_{i \ldots j}$ is unpaired, i.e. accessible, for all folding windows which contain $R_{i \ldots j}$ (see Section 2.5.6). $\overline{pu}(i, j)$ represents the normalised number of times we expect to observe that the subsequence $R_{i \ldots j}$ is accessible over all windows.

Using the above setup, the accessibilities for each u-region could be assessed independently, with increasing distance from the MRE sites. However, since the flanking mRNA sequences are only aligned by the 3' terminus of the MRE site (position zero) and not by sequence conservation (which generally does not exist in the flanking sequences), assessing differences in these small u-regions would not allow for positional variations of flanking signals. Therefore, as illustrated in Figure 8.1.A, we applied a sliding-window approach using windows of size 20 nt, which is twice as large as U and thus larger than all u values for accessibilities calculated by `RNAplfold`. We define a window W_k^l as the subsequence $R_{k \ldots (k+l-1)}$ that starts at position k in R and has the length l (we used $l = 20$). Then we calculate the mean accessibility in W_k^l for one $u \in \{1, \ldots, U\}$ as

$$\frac{1}{l - u + 1} \sum_{\substack{k \leq i \leq j < k+l, \\ j-i+1=u}} \overline{pu}(i, j),$$

[1] Please note that this experiment was performed prior to the work done in Chapter 6 where we determined better W and L parameter values for `RNAplfold`. Therefore, values for W and L parameters were optimised for best significance results.

the maximum accessibility as

$$\max_{\substack{k \leq i \leq j < k+l, \\ j-i+1=u}} \{\overline{pu}(i,j)\},$$

and the minimum accessiblity analogously to the maximum. In Figure 8.1.A, we abbreviated these equations to $mean\{\overline{pu}(i,j)\}$, $max\{\overline{pu}(i,j)\}, min\{\overline{pu}(i,j)\}$ for the mean, maximum and minimum accessibility calculations per window for a fixed u, respectively. For each window sequence W_k^l, the distributions of mean, maximum, and minimum accessibilities were calculated for all sequences in the dataset. Subsequently, sequences containing functional MREs were compared with sequences containing non-functional MREs. A separate comparison is performed for each $u \in \{1, \ldots, U\}$. To assess whether the distributions per window were significantly different, we performed both a two-sample Student's t-test and a two-sample Wilcoxon Rank Sum test; the latter test is non-parametric and thus independent of the type of distribution. The results were plotted in Figure 8.1.B where the t-values are represented by dots and if $p \leq 0.5$ was achieved for the Wilcoxon test the t-value is plotted using a solid line (otherwise the solid line is at zero). it is indicated by a solid line for t-values that is not zero. Thus, results from both tests were combined in a single visualisation.

8.2.2 Higher accessibilities were observed 20 to 50 nt downstream of MREs

In Figure 8.1.B, we observed a clear enrichment for windows with higher accessibilities for functional MRE sites in comparison with the non-functional set. The region of significant accessibilities are approximately between 20 to 50 nt downstream of the MRE sites; the signal is robust for all mean, minimum and maximum accessibility calculations. This region is noticeably separated from the MRE site, alluding to a potential independent recognition motif. Moreover, there is no discernible signal for higher accessibilities overlapping with the MRE site. In addition, disjoint regions of higher accessibilities not overlapping with MRE sites were also observed for two independent datasets from human and firefly, despite these being animals and harbouring distinct differences in their RNAi pathways in comparison with plants [6]. The same results were also observed for siRNA binding sites (siRNAs are one of the other classes of small ncRNA commonly integrated into Argonaute proteins in RISC complexes, see Section 2.2). See the results for the independent datasets in Section D.4.

8.3 Nucleotide frequencies corroborate accessibility signals

In addition to signals of higher accessibilities around MRE sites, we performed a simple analysis of nucleotide compositions in the same windows as in Section 8.2. Instead of calculating mean, minimum, and maximum accessibilities per window, we calculated the frequencies of single nucleotides, of $G + C$, $G + U$ and $C + U$ (other combinations are given by the inverse of the three given), and of all dinucleotides and the significant results were

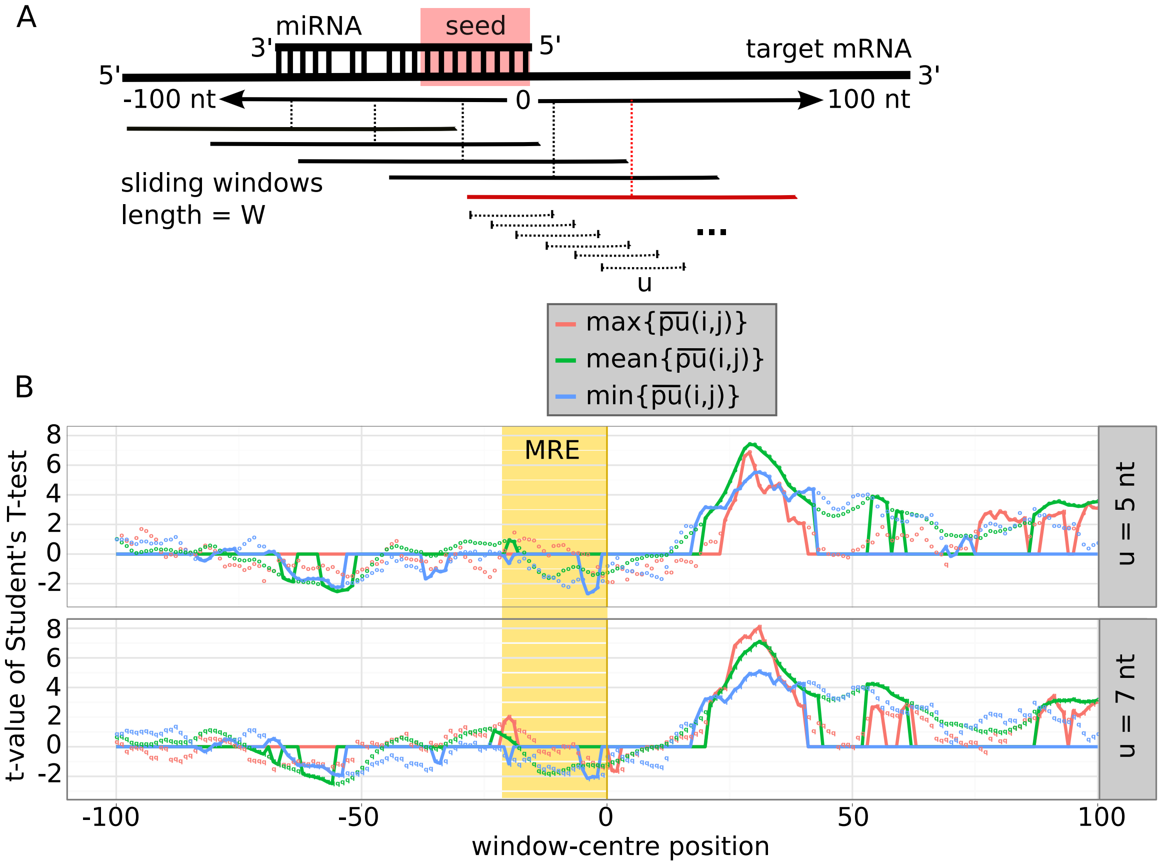


Figure 8.1. Accessibility is significantly higher downstream of MRE sites in *A. thaliana*. Extended context sequences surrounding MRE sites are aligned by the position that is matching to the first nucleotide in the miRNA by setting this to position zero. (A) In a sliding-window approach, the accessibility of each window is measured using `RNAplfold` predictions for regions $u = 5$ and $u = 7$. (B) The centre position of a window is plotted on the x-axis and the t-value of a Student's t-test comparing the distributions of accessibility measurements between the 110 functional and 114 non-functional MRE sites on the y-axis. When $p \leq 0.05$ was achieved for an independent non-parametric Wilcoxon Rank Sum test on the same data, the t-value was plotted using a solid line, otherwise the solid line is at zero. No correction for multiple testing was performed here.

plotted in Figure 8.2. The most notable result is that there is a clear enrichment of C and U nucleotides and UC and CC dinucleotides from approximately 0 to 50 nt downstream of MREs. Correspondingly, the nucleotides G and A and dinucleotides AG and GA are depleted in the same region.

8.4 Conclusion

We observed an enrichment of structural accessibility and a high C and U single- and dinucleotide compositions in a region flanking the MRE sites of *A. thaliana*. Of special interest is that this signal did not occur within the MRE site, but instead 0 to 50 nt downstream. This observation suggests a further recognition element that is not the MRE and could possibly be indicative of Argonaute binding or the binding site of another cooperating factor. In a recent publication, it was shown that Argonaute may bind to an independent motif with increased accessibility 10–20 nt upstream of the MRE seed site in humans that was A rich [192]. This motif was not validated with the presented approach in *A. thaliana.* Instead we see the signal downstream and a depletion in As. In both work, however, there is a signal independent of the MRE site. We further support this by identifying similar independent signals if higher accessibilities in human and firefly MRE sites, and also for siRNA data.

From the data presented in this Chapter, it is difficult to deduce, whether accessibility or an affinity to C and U nucleotides is more important downstream of MRE sites. Therefore, additional analyses, especially wet-lab experiments are required to fully characterise the reason for the observed signals. Nonetheless, it is clear that the context region surround MRE sites contain significant signals of sequence and structure. Identifying independent recognition sites in the vicinity of MRE sites could possibly improve the computational detection of miRNA targets.

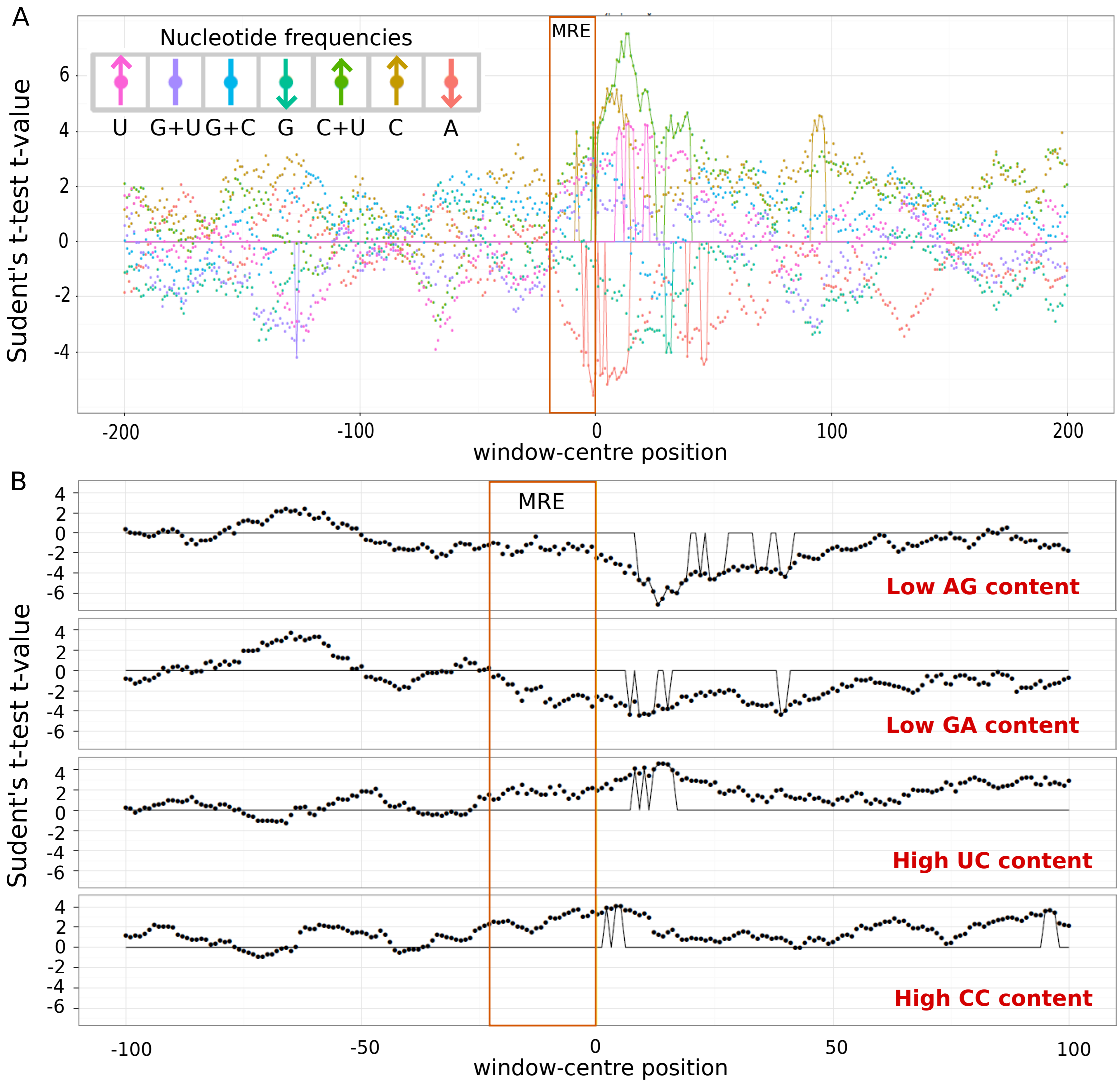


Figure 8.2. Significant signal of nucleotide composition downstream of MRE sites in *A. thaliana*. We assessed the significance of enriched or depleted nucleotide compositions in (A) and dinucleotide frequencies in (B) by calculating the respective values for sliding windows of 20 nt. Each time, we compared distributions calculated from the functional and non-functional MRE sites described in Section 8.1. The centre position of a window is plotted on the x-axis and the t-value of a Student's t-test on the y-axis. When the p-value for an an independent non-parametric Wilcoxon Rank Sum test was significant with , then the t-values are indicated by a solid line; otherwise the solid line is at zero. The significance threshold was $p \leq 0.05$, applying the Bonferroni correction for multiple testing.

CHAPTER 9

A framework for modelling regulatory recognition elements

Accurate *in-silico* detection of RREs from a plethora of *trans* factors that regulate gene expression has remained a troublesome and elusive goal in cell regulatory research—despite a desperate demand for efficient computational tools. Currently, very little is known about characteristics that define interactions between *trans* factors and their respective RREs. In the past, the major bottleneck has been a lack of data, i.e., large, high fidelity datasets of functional and non-functional interaction candidates that could be used for learning interaction models were rare or mostly unavailable. The recent application of diverse `CLIP-seq` protocols to elucidate interaction sites in specific tissues or cells has produced a welcomed influx of data (see Section 2.7.1). We developed a dynamic and flexible machine-learning framework that exploits this new source of data to capture sequence and structure binding affinities of *trans* factors and to support the computational detection of RREs. Trained models can be applied to any other cell line or tissue to detect further *trans*-factor–RRE interactions. Biological insights into characteristics of such interactions can evolve rapidly and thus computational tools must adapt at the same speed. The special advantage of the herein presented framework is that proposed graph encoding of interaction sites can easily and efficiently be modified and extended to include additional knowledge or hypotheses as they become available.

The proposed machine-learning framework can be adapted to model binding preferences of any type of *trans*-factor–RRE interaction. Its application to RBP-RRE interactions was already very successful and has been published under the name `GraphProt` in [P6]. `GraphProt` was mainly developed by Daniel Maticzka and will be part of his dissertation, therefore, to limit overlap, the focus of this chapter is its extension and application to miRNA-MRE interactions (in humans), which should be viewed as work in progress.

9.1 A natural encoding of regulatory recognition elements

Conventional approaches for detecting miRNA target genes learn prediction models from a handful of pre-calculated features about the miRNA-MRE interactions (Section 2.7.3), e.g., the number of base pairs within the seed interaction or the extended hybrid, accessibility of the target mRNA within the MRE and of the surrounding sequence, and hybridisation free energies. In this work, we propose an encoding of miRNA-MRE interactions that is more natural. The key idea is to simply encode the interaction as a graph, which is subsequently processed into thousands, or millions of features[1] that enable a robust comparison between functional and non-functional interactions; a process that is reminiscent of using k-mer frequencies to compare the similarity of strings, e.g., huge molecular sequences [62].

9.1.1 General graph encoding of any regulatory recognition element

The foremost property for *trans*-factor binding specificity is the nucleotide sequence [249]. For miRNAs, the first characteristic to be identified was the seed interaction [11,166,190,191,285], and recent evidence suggests that human Argonautes bind to an A-rich motif, 10–20 nt upstream of the seed interaction [192]. To capture sequence preferences, the primary structure of an RRE can be represented as a simple (chain) graph, connecting the nucleotides according to their backbone structure. This simple model has already shown high performances for many RBPs [P6]. There are many *trans* factors, however, where not only sequence affinity is important, but structural affinity is also a discerning factor. For example, the A-rich motif for Argonaute was observed to be structurally accessible [192], and some Cas6 proteins bind specifically to short hairpin structures (Part II). We can extend the simple sequence model to include the structural context of RREs as illustrated in Figure 9.1. First, the RRE is extracted from its endogenous sequence with sufficient sequence context to calculate accurate local structures[2]. Second, since we require fixed structures, we generate several alternate, probable folding hypotheses using `RNAshapes` [296]; in this way, we avoid the error-prone use of only the MFE structure (Section 2.5). `RNAshapes` categorises the ensemble of all possible structures into several shape-abstraction classes, called *shapes*: the MFE structure within each shape class is called the *shrep*. We use the shrep structure for each probable shape class for our structure encoding. Both the number of shreps/shapes chosen for the encoding (three for this work) and the shape-abstraction level (also three) are parameters of the encoding. Stacking base pairs are further highlighted by an additional vertex with edges between it and the four nucleotides involved (not depicted in Figure 9.1 for clarity). Third, we can extend the basic, secondary-structure graphs (with nucleotides and base pairing information) to hypergraphs that annotate for each nucleotide, the secondary structure element (Section 2.4.1) it belongs to. Hence, sequence, base pairing, and structure elements are modelled together. Although a single graph is given as the input encoding, graphs that

[1] Both the number of features determined by the model hyperparameters R and D and the weights learned for the features are regularised so that overfitting to the training data does not become a problem.

[2] According to the results in Chapter 6, about 100–150 nt on either side of the RRE would be sufficient, we use 150 nt here.

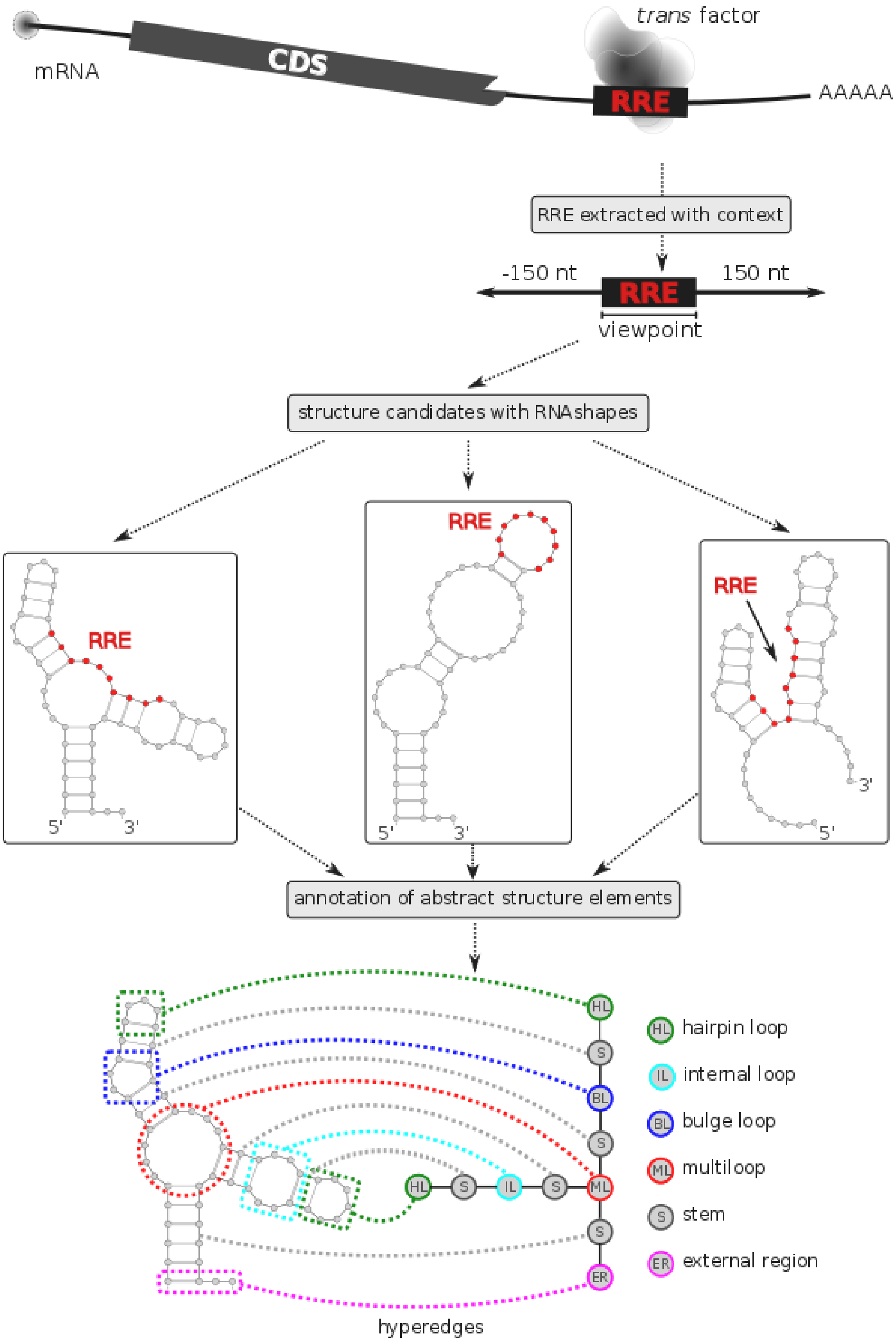


Figure 9.1. A natural encoding of regulatory recognition elements as graphs. First, a subsequence is extracted from the mRNA where the RRE is embedded within its natural sequence context; the RRE site is marked internally as a *viewpoint*. Second, multiple folding hypotheses are generated for the extracted subsequence using `RNAshapes`; structures are represented in the graph format with vertices labelled with the respective nucleotides. Third, each structure candidate graph is extended to a hypergraph to also encode the abstract RNA shape in the form of secondary structure elements (see Section 2.4.1, Definition 2.11). Nucleotides from the base level are linked to their respective element in the abstract level of the hypergraph via hyperedges.

encode increasing levels of information are encoded in separate, disjoint subgraphs. For example, the sequence-only graph is combined with the full structure graph (or hypergraph) as unconnected subgraphs. In comparison with the sequence-only model, the full structure model delivered increased performances for selected RBPs where it is assumed that structure plays an additional role in identifying target RREs [P6].

9.1.2 Extension to miRNA recognition elements

To avoid having to select exact features of the hybrid pattern between miRNA and MRE, e.g., the number of base pairs within the seed region, or the size of bulges, etc., we extended the general RRE encoding by simply adding the mature miRNA sequence and the intermolecular base pairs (see Figure 9.2). The pairs of subgraphs subsequently extracted by the graph kernel capture both the overall structure of the hybrid and whether the MRE is structurally accessible or not.

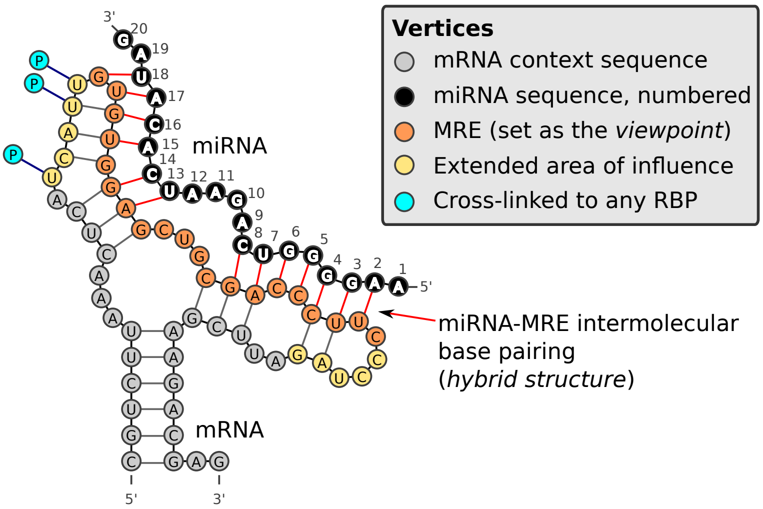

Vertices
○ mRNA context sequence
● miRNA sequence, numbered
● MRE (set as the *viewpoint*)
○ Extended area of influence
● Cross-linked to any RBP
miRNA
mRNA
miRNA-MRE intermolecular base pairing (*hybrid structure*)
3'
5'
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
P

Figure 9.2. Full encoding of miRNA-MRE interactions. The mature miRNA sequence is added to the mRNA local structure with the intermolecular base pairs within the hybrid structure. The MRE is the region on the mRNA that is covered by the miRNA, starting from the first and ending with the last nucleotide. The MRE is set as the *viewpoint* and the added area of influence around the viewpoint due to the feature extraction process is also indicated in yellow. Nucleotides that display at least two T to C conversions (U to C in the RNA) in the RBP-binding profiles from [12] are extended by a further vertex (blue), linked by a single edge.

Recently, an effort was made to catalogue all RBPs that bind to mRNAs in the human HEK293 cells [12, 38, 214]. Instead of purifying a selected cross-linked protein as in the usual `CLIP-seq` approach, these authors have modified the protocol to purify all RNA that has been cross-linked to *any* RBP. In such a way, transcriptome-wide signatures of any RBP binding event can be measured. In the literature, there is both evidence of cooperative RBPs that can enhance miRNA regulatory effects (e.g., the additional binding of the Argonaute protein [192] or binding of Pumilio proteins [148, 162]) or sequester miRNA binding (e.g., as has been observed for HuR [42]). To capture possible cooperative or competitive effects of RBPs we have extracted the data from [12] (see Section D.4 for details) and have added additional vertices to U nucleotide-vertices that display at least two T to C mutations in the RBP-binding profiles.

The full encoding of the miRNA-MRE interaction does not have to be used. We explored various encodings, starting from the most basic representation, and extending this hierarchically to the full model. All explored models are summarised in Table 9.1.

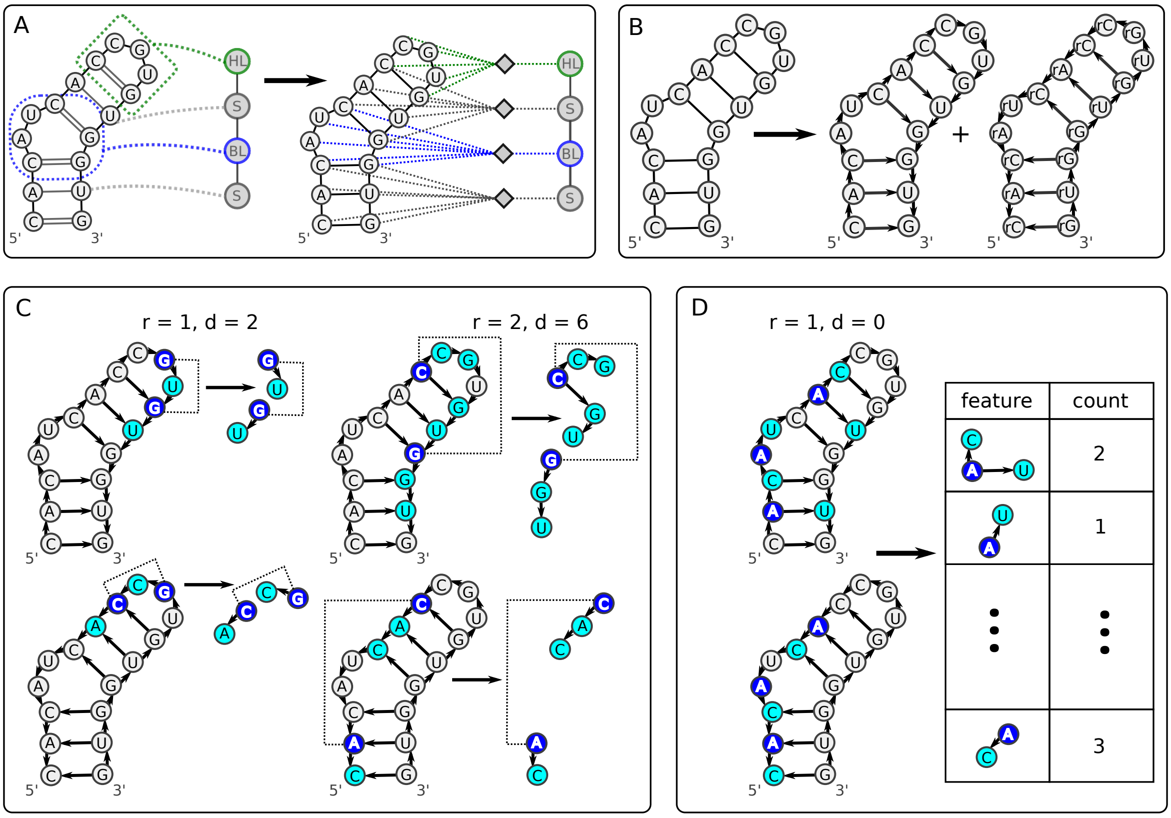
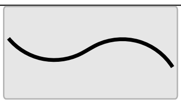
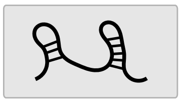
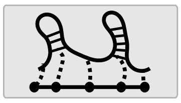
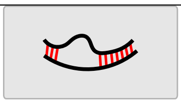
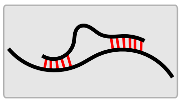
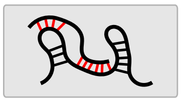
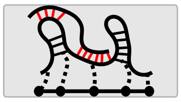
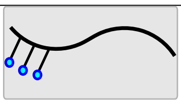
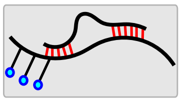
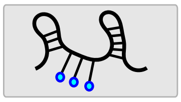
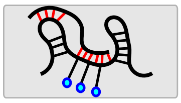


Figure 9.3. Feature extraction of the NSPD Kernel. (A) The hypergraph encoding abstract structure elements is transformed into the respective incident graph by adding additional vertices for each hyperedge relation that links the basic structure with its structure element. (B) The undirected graph is converted to a directed graph to reflect the natural 5' to 3' orientation of RNA molecules; to retain all information, the graph is duplicated with the copy containing inverted edge directions and modified labelling. (C) The graph kernel decomposes the input graph into pairs of neighbourhood subgraphs. Each subgraph pair is determined by a radius $r \leq R$ and a distance $d \leq D$ where R and D are parameters set by the user. The dark blue vertices are roots from which the path lengths are determined. (D) The frequencies of each and every pair of subgraphs are stored in a feature vector.

9.2 Feature extraction and model building

Once functional and non-functional miRNA-MRE interations have been encoded as one of the (hyper)graphs in Table 9.1, we apply a machine-learning technique to differentiate between the two sets. Conventional machine learning approaches require a feature vector to describe each instance. For this purpose, we apply a graph kernel to convert the input graph into feature vectors (see Figure 9.3). Subsequent to the conversion, any machine learning approach can be applied to the feature vectors to build prediction models. In this work, we

Table 9.1. Eleven encoding models to capture characteristics of miRNA-MRE interactions. Models 1–3 can be applied to any RRE; models 3–7 are extended to encode mir-MRE sites; models 8–11 add further experimental RBP-binding information.

No.	Model	Name	Description
1		sequence	Encodes the MRE sequence within its mRNA sequence context
2		structure	Encodes the MRE within its mRNA sequence context as secondary structure candidates calculated by `RNAshapes`
3		structure &elements	Extends the basic structure model by annotating abstract structure elements
4		hybrid	Encodes the miRNA and MRE sequences with the intermolecular base pairs as predicted by `IntaRNA`
5		sequence &hybrid	Combines the hybrid graph with the extended context sequence
6		structure &hybrid	Extends the basic structure model with the miRNA hybrid structure
7		structure, hybrid &elements	Extends the structure and abstract structure element graph with the miRNA hybrid information
8		sequence &protein profiles	Extends the sequence-only model with nucleotide-wise graph extensions when there is evidence of RBP binding (blue vertices)
9		sequence,hybrid &protein profiles	Extends the sequence and protein profile model with the miRNA hybrid information
10		structure &protein profiles	Encodes the structure model with added information when there is evidence of RBP binding
11		structure, hybrid &protein profiles	Extends the structure and hybrid model with added information when there is evidence of RBP binding

applied the Support Vector Machine (SVM) to the present classification task. However, when affinity measures are available, any regression technique can be applied—as was done in [P6]. Prediction performances between various models were compared by the traditional ten-fold cross validation, where a model is first trained on 90 % of the data and then tested on the remaining 10 % of the data, and train–test iterations are performed ten times in total where each of the test datasets do not overlap; the average performance is reported. The following sections briefly describe the graph kernel and extensions to the input graph that are required for extracting accurate features.

9.2.1 Graph kernel

We employed the *Neighborhood Subgraph Pairwise Distance kernel* (NSPD Kernel) [56] to convert input graphs into feature vectors. The main idea of the approach is to decompose the graph into (usually) thousands of small overlapping subgraphs and the final feature vector is represented by sparse vector of subgraph frequencies (illustrated by Figure 9.3.C–D). Every subgraph is assigned a numerical identifier via an efficient hash-based technique. Comparisons on numerical identifiers is not only extremely efficient, but allows a fast (albeit approximate) solution to handling graph isomorphisms. In this way, we can effectively process millions of features that can correspond to large input graphs. In detail, the NSPD Kernel describes features as a conjunction between two neighbourhood subgraphs at a small distance from each other. Two parameters determine the characteristics of these subgraph pairs (and are related to the complexity and size of the entire feature set): (1) the maximum size of the neighbourhood, called the *radius* (R), and (2) the maximum distance between any two root vertices, called the *distance* (D). Each subgraph contains a root vertex from which the radius and distance are counted and features are extracted for all combinations of values $r \leq R$ and $d \leq D$ (Figure 9.3.D for an illustration). Optimal values for R and D change for different models and encodings (see [P6]). Optimising these parameters can therefore increase prediction performances.

To handle the specific demands of comparing RREs, the original NSPD Kernel was extended as follows: (1) simple graph encodings were upgraded to hypergraphs to handle the abstract-structure-element annotations; (2) directed graphs were considered rather than undirected graphs so that only features are considered that regard the 5' to 3' direction of nucleic acids; and (3) to avoid an influx of uninformative features due to the increased sequence context required for accurate structure predictions, we restrict feature extraction to only the RRE region by labelling this region as a *viewpoint*. This means that the kernel extracts informative features by considering only those vertices labelled as viewpoints for the feature extraction process.

A kernel for hypergraphs

The hypergraph that encodes the abstract structure elements is first converted into an incident graph (Figure 9.3). For each one-to-many relation (i.e., hyperedge) between the basic and the abstract structure levels in the hypergraph, a relation vertex is added, with a single edge leading to the abstract structure element and many edges leading to each vertex (i.e. nucleotide) involved in that element. In the NSPD Kernel, published in [56], the graph is decomposed into features (pairs of subgraphs) with respect to the radius R and distance D using shortest paths (see Figure 9.3.C). A problem arises when the graph contains vertices with a large degree (i.e. many connecting edges), as is the case for the hypedges in Figure 9.3.A. In this case, the shortest path distance notion for the feature decomposition becomes degenerate: many vertices become immediate neighbours of each other and the decomposition would result in uninformative features corresponding to extremely large subgraphs. Such large subgraphs are unlikely to occur in more than one instance and this would make effective learning or generalisation impossible. This situation occurs if the incident graph in Figure 9.3.A is used for the hypergraphs: hyperedges (i.e., relations between the basic and the abstract structure levels) yield vertices with a large degree, e.g., a stem relation vertex is connected to all stacking base pairs within the stem. This effectively removes the nucleotide order of the RNA sequence, since there exists a shortest path of length two between any two nucleotides involved in the stem structure. In order to circumvent this problem, the NSPD Kernel was extended to work on the incident graph by (1) considering the relation vertices as non-traversable by paths; and (2) by creating additional pairs of subgraph decompositions where the root vertices of the two paired neighbourhoods are on the two endpoints of the hyperedge relation. This yields features that are aware of the nucleotide composition of a substructure and, at the same time, of the position of that substructure in the global abstract structure. Finally, the updated NSPD Kernel generates three sets of features: one set only describing the basic structure level, a second set only describing the abstract structure, and the third set of features represent relations between the basic and abstract levels.

Directed graphs

To introduce the asymmetry imposed by the 5' to 3' orientation of RNA, we converted the undirected graphs into directed graphs (Figure 9.3.B). To be able to capture all relevant information, while still maintaining consistency with the RNA direction, we duplicated the graph, relabelled all vertices by adding a distinguishing prefix, and reversed the direction of all edges. The NSPD Kernel only traverses paths according to edge directions.

Selection of kernel viewpoints

The selection of kernel *viewpoints* limits the extraction of features to only relevant regions of the input graph. Applying viewpoints allows the use of extended context sequences for accurate structure predictions. Subsequently, only informative features are extracted. Without the viewpoint notion, thousands of uninformative features would be generated, leading to lower prediction performance. More precisely, when a viewpoint is set, at least one of the root vertices in an extracted pair of subgraphs is required to be part of the viewpoint. In this work, we set the viewpoint to cover the entire MRE region (see Figures 9.1 and 9.2). Extracted features do not only include vertices and edges within the viewpoint (i.e. the MRE), but due to the radius (R) and diameter (D) parameters, an extended area of influence exists that reaches beyond the MRE (Figure 9.2); this area of influence extends to a maximal distance of $R + D$ from vertices within the viewpoint. Setting viewpoints that were symmetrically larger than the MRE did not increase prediction performances (data not shown). The viewpoint technique was first introduced in [90].

9.3 Application to miRNA recognition elements

To test whether the proposed models can accurately detect miRNA-MRE interactions, we first required a large set of exact interaction data. No sufficient dataset exists in the literature, therefore, we took careful measures to curate a suitable dataset from recent `CLIP-seq` experiments. Although these experiments detect Argonaute-RRE interactions, they usually do not give any insight into which miRNA is involved—if any is involved at all. Hence, we applied extensive measures to procure high-quality hybrid structure predictions between the identified Argonaute-RRE and the best-matching expressed miRNA sequence.

In this work, we focus on deriving the best model for encoding miRNA-MRE interactions. For the first time, we generate single-miRNA models to encode specific preferences of individual miRNAs. In addition, we determine the generalisation capabilities of trained models to previously unseen miRNAs.

9.3.1 Acquisition of high-fidelity miRNA-MRE interaction data

Models were trained and tested on carefully curated functional and non-functional miRNA-MRE interactions that were derived from `CLIP-seq` experiments [122,135,172] with cross-linking to Argonaute proteins in HEK293 cells. Using more than 14,000 filtered AGO1–4 RRE sequences, we selected functional sets of interactions for the miRNAs with the highest expression levels in the same HEK293 cells. Corresponding non-functional sites were selected from transcripts that contained—but did not overlap with—cross-linked sites from all available `CLIP-seq` experiments [122,135,172] (see Section D.4 for details). Both functional and non-functional sets contained roughly the same number of—and for each miRNA, at least a few hundred—interactions. Our extensive efforts ensured that the functional interactions closely resembled the non-functional interactions. Both sets contained the same type of seed interactions so that discerning the difference between the two classes was extremely difficult. In addition, to select anchors for calculating candidate miRNA-MRE hybrids, we first scanned potential target sequences for the locations of various seed types that have been observed to cause a regulatory effect in experiments reported in the literature [45,65,121,196,216,306,306]. The complex filtering and processing steps that were necessary to generate the datasets are described in Section D.4.

9.3.2 Performance comparisons of various encoding models

We first trained and tested all encoding models from Table 9.1 on five selected miRNAs individually. For each miRNA-model combination, the parameter settings for the graph, maximum radius R and distance D, were optimised, testing values 1 to 4 and 1 to 6, respectively. Performance measures were reported as the area under the receiver operating characteristic (AUROC) in a ten-fold cross validation setting in Table 9.2.

Although all encoding models displayed a certain degree of predictive power (any AUROC value above 0.7 is an acceptable performance), we observed variations between the different models (Table 9.2) that were consistent for each of the tested miRNAs. The first significant result is that in this setting, structure information did not improve prediction performance. It is possible that the current structure encoding does not capture the relevant information for miRNA-MRE interactions. However, the finding is consistent with results from Chapter 8, in which we determined that the actual MRE site does not display a significant signature of higher accessible structures. The accessible region downstream of MRE sites could be too far away in humans to be able to capture this in current models. Extending the viewpoint beyond the MRE did not improve performances due to the massive increase in uninformative features, therefore, one would have to further narrow down the second area of influence to possibly increase predictive power.

Overall, even the simple sequence model (model 1 in Table 9.2) is powerful, achieving an average AUROC of 0.84 for the five miRNAs. Adding the miRNA hybrid structure to the MRE sequence does increase the performance slightly; notably adding extended context information around the MRE site performs better than just restricting the encoding to the miRNA-MRE hybrid structure. A more marked increase in performance is achieved by adding the RBP binding information (model 9 in Table 9.2). Whether the signal that is captured is Argonaute binding or the binding of a different RBP remains to be seen.

In summary, the best model for detecting miRNA-MRE interactions is one that includes the context sequence of the MRE, the hybrid structure with the miRNA and RBP-binding profiles, if available.

9.3.3 The ability of trained models to predict interactions for miRNA not in the training data

To test the generalisation capabilities of the two best-performing encoding models, we set up a special ten-fold cross validation task where we test on interactions for miRNA that were "unseen" in the training phase. In each iteration, we trained on data containing interactions for nine miRNAs and tested on the interaction data for the one miRNA that was excluded from the training data. This was iterated ten times where each of the ten miRNAs was used for training once. For this experiment, the ten miRNAs with the top expression levels were extracted from the AGO1–4 `PAR-CLIP` curated dataset, each with hundreds of functional

Table 9.2. Comparison of different encoding model for interaction data comprising of single miRNAs. Parameters R and D are optimised for each miRNA–model combination. Interaction data for the five tested miRNAs[1] are extracted from the AGO1–4 `PAR-CLIP` curated dataset (Section D.4); in addition, interactions were filtered to take only the top 50 % with the highest read coverage for the RREs detected in the `PAR-CLIP` experiment [122]. The performance is measured as the AUROC in a ten-fold cross validation setting. Each dataset contained hundreds of balanced functional and non-functional interactions. Two encoding models are selected for subsequent use that performed best and belong to different categories of information (in bold).

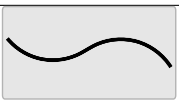
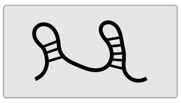
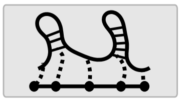
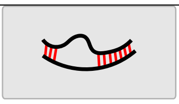
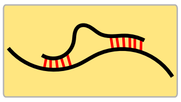
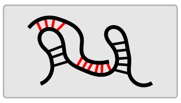
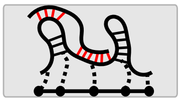
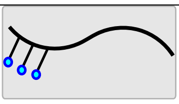
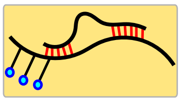
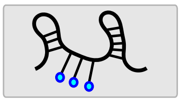
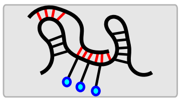
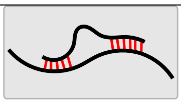
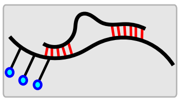
No.	Model	Name	mi1	mi2	mi3	mi4	mi5	avg.
1		sequence	0.83	0.81	0.84	0.86	0.84	0.84
2		structure	0.79	0.75	0.78	0.83	0.80	0.79
3		structure & abstract shape	0.71	0.67	0.70	0.73	0.74	0.71
4		hybrid	0.85	0.78	0.83	0.87	0.85	0.84
5		**sequence & hybrid**	**0.87**	**0.81**	**0.85**	**0.87**	**0.87**	**0.85**
6		structure & hybrid	0.84	0.76	0.82	0.86	0.84	0.82
7		structure, hybrid & abstract shape	0.80	0.70	0.77	0.80	0.81	0.78
8		sequence & protein profiles	0.86	0.84	0.84	0.89	0.89	0.86
9		**sequence, hybrid & protein profiles**	**0.88**	**0.85**	**0.86**	**0.90**	**0.91**	**0.88**
10		structure & protein profiles	0.82	0.81	0.80	0.86	0.85	0.83
11		structure, hybrid & protein profiles	0.88	0.83	0.84	0.88	0.89	0.86

Table 9.3. Testing the generalisation capability of trained models to unseen miRNA. In a leave-one-miRNA-out cross validation setting on the ten highest expressed miRNA in the AGO1–4 `PAR-CLIP` curated dataset, we assessed the ability of models to predict miRNAs previously unseen in the training phase. The average performance was reported for the ten iterations. The same R and D parameters were applied that were optimal in Table 9.2 for non-overlapping interaction data.

No.	Model	Sensitivity	Specificity	Precision	AUROC
5		0.68	0.69	0.69	0.75
9		0.78	0.68	0.71	0.81

and non-functional interaction sets[1]. It is important to note that the seed sequences for all ten miRNAs differ enough that models are not just capturing seed-complementary regions. The AUROC results of more than 0.75 in Table 9.3 give an indication that it is possible to learn from interaction data including a mixed set of miRNAs and use this model to detect MRE sites for previously unseen miRNAs, i.e., miRNAs not in the training set. However, the performance did drop in comparison with models using only single-miRNA interaction data. Although it is easier to learn from single miRNAs, this is not very suitable for practical prediction approaches. First, one would have to build a model for each miRNA separately, and there exist over 20 thousand miRNA genes in the `miRBase` [176]. Second, even if we could create a model for each miRNA, it is unlikely that enough data will exist for training for every miRNA; especially for miRNAs with very few targets. Thus, the generalisation capability of trained models is vital to the success of the proposed method.

9.4 Conclusion and outlook

We have proposed a flexible and efficient graph kernel for RNAs that is capable of capturing binding characteristics of any RRE and specifically MREs. The key flexibility of the approach is that one only has to change the input graph encoding that can generally be processed by the NSPD Kernel into feature vectors used by any machine learning approach. Comparing instances via the frequencies of thousands of tiny subgraphs is comparable to comparing strings using the frequencies of k-mers; this approach is robust and it is unnecessary to perform additional feature selection as is done in conventional machine-learning approaches [43].

In this work, we determined that the best way to encode miRNA-MRE interactions (for our data) is a graph that includes the MRE, a small area of influencing context sequence, and the hybridisation structure between miRNA and MRE. An additional boost in performance can be achieved by exploiting the availability of transcriptome-wide binding profiles of arbitrary

[1] The `miRBase` [176] identifiers for the ten selected mature miRNAs were as follows: hsa-miR-30e-5p, hsa-miR-103a-3p, hsa-miR-21-5p, hsa-miR-423-3p, hsa-miR-92a-3p, hsa-miR-19b-3p, hsa-miR-10a-5p, hsa-miR-let-7a-5p, hsa-miR-301a-3p, hsa-miR-93-5p.

RBPs [12]. Models that were trained on miRNA–interaction data comprising of only a single miRNA resulted in the best performances, however, when testing on multiple-miRNA-interaction data, acceptable performances were also achieved. Hence it should be possible to create general models that can predict target sites for any miRNA.

We have already published results that prove the exceptional ability of the RNA graph kernel to capture RBP binding characteristics. Although we observed a decrease in performance when using the structure model for miRNAs, selected RBPs displayed an increased performance in comparison with the sequence-only model [P6]. Furthermore, the RNA-sequence-and-structure encoding described in Section 9.1.1 was successfully applied in an independent study to the clustering of thousands of RNA sequences to detect conserved families of functional non-coding RNA [136]. Taken together, these results are indicative of a very promising application of the RNA graph kernel to predict MREs for any miRNA. Although to prove its suitability, extensive benchmarks with state-of-the-art prediction approaches are still outstanding. Comparisons are currently being processed. Such benchmarks of miRNA prediction tools are an extremely time-consuming task: many approaches do not provide software that is suitable for high-throughput predictions and their published results are difficult to map to the data used in this work due to different genome assemblies and missing or incompatible sequence identifiers. Moreover, we are planning to extend the generalisation capability of the models and curate improved interaction data.

We are currently working on a procedure to extract information about informative features from the trained models. Since we model the input as a graph that represents the natural setting of a miRNA-MRE interaction, informative features can be mapped back onto the input graph to understand what the model has learnt. A simple approach to extract sequence and structure profiles was already implemented for `GraphProt` and successfully identified information about RBP affinities [P6], however, profiles reduce the complete information available in the full graph encodings.

A general limitation of the presented approach is that sufficient training material is required for generating prediction models. Therefore, future effort should be put into determining the capability of models trained on one species to predict miRNA interactions in closely-related species. For example, if the models presented in this work, trained on human data, could predict interactions occurring in mice, this would greatly increase the scope and applicability of RNA graph kernels.

Part VI

Targeted gene regulation using artificial RNA

Part VI: Targeted gene regulation using artificial RNA

Knowledge is not wisdom, unless used wisely.—J. D. Anderson

Research of natural sciences is roughly divided into two major goals: (1) learning more about fundamental concepts, processes and mechanisms that make up what we call *nature*; and (2) using the knowledge gained to 'improve' our lifestyle by curing diseases, developing new technologies, creating more efficient processes, or genetically modifying organisms to suit some requirement conceived by humans. Artificially designed, RNA-guided, regulatory mechanisms can provide a powerful tool when striving to achieve either goal. The endogenous CRISPR-Cas and miRNA-based regulatory systems, both frequent topics throughout this thesis, have been successfully adapted to the task of artificially silencing the expression of any gene of interest: either on the level of transcription by targeting DNA [328], or post-transcriptionally by targeting RNA and inhibiting protein production [167, 236, 331]. When the expression level of a gene of interest has been altered, its effect on other processes can be deduced to gain insight into its function in the cell. Once a favourable effect has been observed, organisms can be permanently or transiently altered to produce that effect. Artificially altering the genetic material of a living organism, outside of controlled laboratory experiments, can have wide-spread effects on other living organisms or on entire ecosystems that are almost impossible to predict or foresee. Therefore, this technology raises many ethical questions and should be handled with great care.

In Parts II–V, we explored different aspects of post-transcriptional regulatory mechanisms where RNA sequence and structure is a key driving force behind post-transcriptional gene regulation. Here, in Chapter 10, we applied knowledge, previously gained from endogenous systems, to the design of artificial regulatory constructs. In particular, we explored characteristics of the interaction site that affected the repression efficacy of artificial miRNAs (amiRNAs) in *Arabidopsis thaliana*, a model organism for plants.

CHAPTER 10

Artificial microRNAs that suppress gene expression in plants

Suppressing gene expression in plants for functional studies is commonly achieved by causing nonsense mutations via T-DNA insertions into the gene of interest [179] or by targeted transcript degradation using RNA interference (RNAi) [1, 353]. T-DNA insertions are not available for all plant genes and they lead to a complete loss of function, which may be lethal in some cases. The advantages of RNAi are that guide RNAs can be designed to target any gene of interest or even multiple genes; it can lead to a finer regulation of the target gene where the expression is decreased enough to analyse its function but is still sufficient for cell survival; and it can be applied so that the suppression is reversible.

Central to RNAi are the guide RNAs that determine target-site specificity: generally siRNAs[1] or miRNAs (see Section 2.2.1). Although siRNAs are commonly used for artificial gene silencing, they frequently lead to the generation of secondary siRNAs that cause the undesirable silencing of off-target genes [83]. In contrast to animal miRNAs, miRNAs in plants display extensive complementarity to their target site: stable and extended base-pairing between miRNA and MRE covering the 10–11th nucleotides of the miRNA (counting from the 5') leads to a cleavage and subsequent degradation of the target transcript [6]. This extended complementarity in plants means that artificial miRNAs (amiRNAs) can be used analogously to siRNAs for specific targeting of single genes; although amiRNAs could still regulate off-target genes, these can be minimised in the design phase.

A mature amiRNA is processed from an endogenous, template pri-miRNA where the miRNA and the miRNA* (the passenger strand; Section 2.2.1) have been replaced by sequences designed specifically for a target of choice. For subsequent analyses, ath-MIR319a (MI0000544, miRBase [112]) was used as the template. The endogenous miRNA processing pathway

[1] In *A. thaliana* there are at least three different types of siRNAs: *trans*-acting siRNAs (ta-siRNAs), repeat-associated siRNAs (ra-siRNAs), and natural antisense transcript siRNAs (nat-siRNAs) [15].

generates the mature amiRNA that is subsequently assembled into a RISC complex[1] to bind and degrade the target mRNA transcript. An amiRNA is 21 nt long and is designed so that most bases are complementary to its target[2]. The corresponding amiRNA* can be derived from the amiRNA such that the exact bulge structure of the pri-mRNA, the stem-loop fold of ath-MIR319a, is maintained; bulges within this region are important for correct processing.

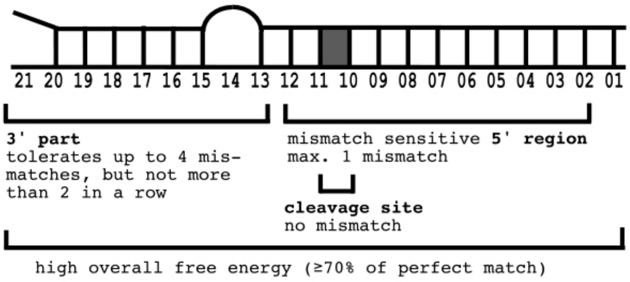


Figure 10.1. WMD amiRNA design features. A schematic illustration of the design features of the predicted amiRNA–target duplex structure for WMD [236,331]. The target site is shown at the top in $5' \rightarrow 3'$ orientation and the amiRNA is below in the opposite orientation; numbering is according to the amiRNA sequence. Figure taken from the WMD website (`http://wmd3.weigelworld.org`).

The Web MicroRNA Designer (WMD; `http://wmd3.weigelworld.org`) provides a popular and easy-to-use web interface for the design of suitable amiRNA sequences for many model plant species [236]. The 'design' tool of WMD proceeds in two main steps: (1) suitable amiRNAs are chosen to reflect rules mimicking mammalian siRNAs [236], and (2) calculation of candidate amiRNA specificity, i.e., the number of off-targets that they are predicted to regulate. The amiRNAs are ranked by a cumulative score and assigned a colour that reflects their predicted functionality: green is very favourable, whereas orange and red indicate amiRNAs with potentially reduced efficiency or specificity, although all reported amiRNAs are deemed functional by the authors. Based on intermolecular base-pairing interactions and hybridisation energies predicted by `RNAcofold` [19], WMD uses the following criteria to design efficient amiRNA [236,331]:

- *A* (or *U* for multiple targets, if required) at position 10 and a *U* at position 1 of the amiRNA.
- 5' instability of the amiRNA with a higher *AU* content at the 5' than at the 3' end.
- The following constraints were used for the duplex structure between amiRNA and predicted target site, since it was suggested by the authors of WMD that perfect complementarity was not as effective for siRNAs: at most one unpaired base from positions 2–12; up to four unpaired bases, but not more than two consecutive unpaired bases, from positions 13–21; and no unpairing is allowed at the cleavage site, position 10–11. The preference is to have no unpairing between positions 2–12, but 1–2 unpaired

[1] MicroRNAs should be active in all cell types, however, the expression levels of genes involved in the biogenesis and targeting should be verified if experiments are failing.

[2] We use the term "target site" for amiRNA instead of MRE because these sites are not native recognition elements, but artificial.

bases between positions 17–21 (Figure 10.1). Positions are given w.r.t. the amiRNA. A pair of unpaired bases in amiRNA and target is referred to as a 'mismatch'.

- The hybridisation energy between amiRNA and potential target must exceed 70 % of its optimal pairing energy (i.e., the hybridisation energy between the amiRNA and its reverse complement sequence); in addition, the energy must exceed −30 kcal/mol^{-1}. Preferred relative hybridisation energies are between 80–95 % with an absolute value between −35 to −38 kcal/mol^{-1}.

In this chapter, we used WMD to design amiRNAs against target genes in *Arabidopsis thaliana*. We investigated features of the hybridisation pattern and of the sequence context surrounding the potential target site that could affect knock-down efficacies.

10.1 Efficiency analysis of WMD-designed amiRNAs

Within the framework of his Ph.D. thesis, Claude Becker and colleagues used the WMD2 framework to design 62 amiRNAs (coloured green by WMD) against PIN1 (*Arabidopsis* information resource (TAIR) identifier AT1G73590.1 [184]). They developed a fast, flexible experimental protocol to test the knock-down efficiency of such a large set of amiRNAs using a vector system in protoplast cells [15]. The amiRNA screening vector consists of (1) the target gene fused (translationally) to a fluorescent reporter, in this case the green fluorescent protein (GFP); (2) a transformation marker, mCherry, that allows the identification of transformed cells; and (3) the designed amiRNA gene. Protoplasts that are successfully transformed with the vector, determined by the presence of mCherry, are analysed under a microscope to measure GFP fluorescence. Since the target gene is fused translationally to the GFP reporter, only a single fusion mRNA transcript is produced. If the amiRNA is active, it degrades the fusion mRNA and a low fluorescence signal is expected; the reverse is true for amiRNAs that are inactive. For each cell in a sample, the mean pixel intensity per cell was reported. Although this method can be sensitive to localisation within the protoplast, PIN1 is expressed in the cytoplasm, so measurements were expected to be acceptable. The designed amiRNAs and corresponding amiRNA* sequences were published in Chapter II of Dr. Claude Becker's dissertation [15]. This thesis also provides all experimental details.

10.1.1 Generating efficacy scores from GFP fluorescence measurements

The first task was to convert the mean pixel intensities of GFP fluorescence across all protoplasts in a sample into a single efficacy score that reflects the knock-down activity of the amiRNA expressed in that sample. In Appendix D.5, we summarise mean fluorescence intensities per protoplast in a single sample as a boxplot (Figure D.22). Here, the major problem of these data is evident: very low intensities are present in all samples and intensities increase steadily until the maximum measurement is reached (not shown). In fact, the main discerning factor between samples is the maximum fluorescence measurement. Possible reasons for no discernible GFP fluorescence—besides amiRNA-induced degradation of the mRNA—could be that the fusion GFP was not active or perhaps not easily detected in the measured layer of focus of the microscope. The consequence of this behaviour is that the application of standard mean or median values is not robust for summarising the data due to high variances; also, using the maximum value would not be robust to outliers. To partially overcome this problem, we removed 50 % of the lowest intensity values per sample. We report the mean, median (equivalent to the 75 percentile for all data) and the standard deviation of the top 50 % of measurements from all samples in the Appendix D.5. We chose the 75 percentile, scaled to single efficacy $\mathcal{Q}$ values in the interval [0,1] for all subsequent analyses:

$$\mathcal{Q}(\alpha) = \frac{I(\alpha) - I(PC)}{I(NC) - I(PC)}, \tag{10.1}$$

where α denotes the sample containing a single amiRNA; $I(\alpha)$ is the 75 percentile fluorescence intensity measurement for that sample; PC is the positive control (a sample that does not express GFP, denoted as wt); and NC is the negative control (a sample containing the endogenous ath-MIR319a miRNA that does not target PIN1, mCherry or GFP, denoted as mock/mir319a). The PC sample expresses no GFP so that the measurements represent general background flourescence and result in low values. Equation 10.1 can also be referred to the *relative response ratio* as it is relative to both PC and NC. A further control was given by GFP-7, which includes a functional amiRNA against GFP, but was not used for the normalisation (Appendix D.5). The normalised efficacy score of zero corresponds to an amiRNA with full repression activity and one corresponds to no repression ability.

10.1.2 The efficacy of many WMD-designed amiRNA is not sufficient

In recently published work [63], Deveson and colleagues designed four amiRNAs to target MYB33/65 in *A. thaliana*. The miRNAs were designed to mimic the native interactions of ath-MIR159a with its targets; ath-MIR159a was used as the expression template for the amiRNAs. Overall, the features of design were very similar to those used by the WMD design tool. Results showed a large variance in amiRNA efficacy and none of the amiRNAs were as efficient as the endogenous ath-MIR159a miRNA. Our data corroborated this variability in efficacy for the 62 amiRNAs targeting PIN1: although all amiRNAs had been designed according to the same criteria, their expression induced a wide range of target gene suppression levels (see Appendix D.5 and Figure D.22). AmiRNAs were assigned identifiers from P1–P62. If we consider an efficacy score <0.3 to represent a functional amiRNA, then only 20 amiRNAs, approx. 30 %, were functional. A further 8 (13 %) amiRNAs showed no activity with scores above 0.7. The remaining amiRNAs displayed partial knock-down activity. Obviously, there is room for improving amiRNA design.

10.1.3 Additional features that describe measured amiRNA efficacies

We performed a thorough investigation of structure- and sequence-related features, chosen for their potential to directly or indirectly influence the amiRNA-target interaction. The features could be classified into the following categories: accessibility of the target site, hybridisation properties of the amiRNA-target interaction, position-specific nucleotide frequencies of the amiRNA and the target site, intramolecular base pairing in the amiRNA precursor, GC content of the target site, and the unfolding energy inherent to the target mRNA in the region of amiRNA binding. The importance of features was determined by computing correlations between feature values and repression efficiencies. Most features were found to not be correlated with the amiRNA efficacy (data not shown). However, we identified further features of the hybridisation pattern between amiRNA and target site that affected amiRNA efficacy. For design by the WMD tool, `RNAcofold` [19] was used to predict the hybrid. The main problem with `RNAcofold` is that it predicts single structures for the two interacting RNA sequences and therefore several classes of interaction types, for example kissing hairpin loops, cannot be detected. In contrast, we used a more recent approach without such limitations, `IntaRNA` [34], which, in addition, includes the accessibility of the target mRNA in its prediction model[1]. Finally, we reduced all information to four binary features that could explain amiRNA efficacy; the first three are features of the prediction hybridisation between amiRNA and target site and the final feature is of the PIN1 sequence surrounding the target site. The features correspond to the following statements and when true, the amiRNA is generally less efficient than an amiRNA for which no statement applies: (1) more than two consecutive unpaired bases at the 3' end of the amiRNA in the predicted hybrid[2]; (2) less than 15 consecutive base pairs in the hybrid; (3) a bulge in the duplex structure; and (4) target sites that overlap with, or are located very close to, polypyrimidine tracts of at least eight consecutive Us and Cs (Figure 10.2). Note that for every amiRNA binding to this UC-rich region, one of the other three unfavourable features (1)–(3) occurred, however, these were the amiRNAs closest to being completely non-functional.

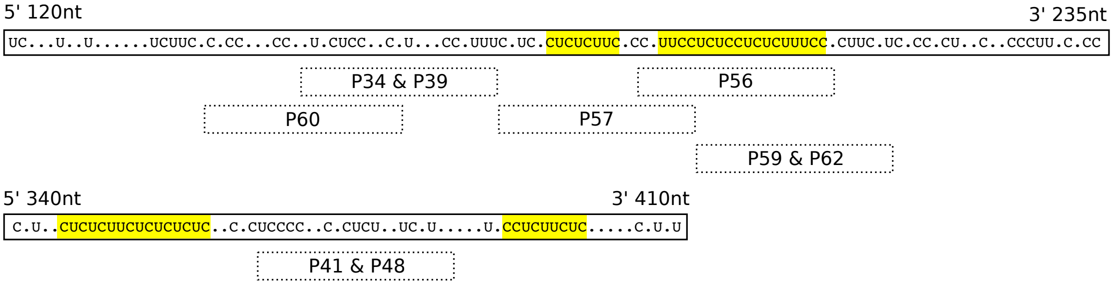

5' 120nt
3' 235nt
UC...U..U......UCUUC.C.CC...CC..U.CUCC..C.U...CC.UUUC.UC.CUCUCUUC.CC.UUCCUCUCCUCUCUUUCC.CUUC.UC.CC.CU..C..CCCUU.C.CC
P34 & P39
P56
P60
P57
P59 & P62
5' 340nt
3' 410nt
C.U..CUCUCUUCUCUCUCUC..C.CUCCCC..C.CUCU..UC.U.....U.CCUCUUCUC.....C.U.U
P41 & P48

Figure 10.2. Polypyrimidine tracts hinder amiRNA repression activity in `A. thaliana`. The target sites of the amiRNAs are indicated below the sequence by the dashed boxes; multiple amiRNAs were designed to interact with the exact same target site. The polypyrimidine tracts of at least eight Us and Cs are highlighted in yellow, As and Gs are replaced by dots to accentuate the UC richness. The indicated amiRNAs display some of the worst efficacy values out of the 62 amiRNAs that were tested: the yellow polypyrimidine tracts are likely blocked by RNA-binding proteins, possibly by the polypyrimidine-tract–binding protein (PTB) [270,275].

[1] `IntaRNA` version 1.2.2 was used with the parameters: `-w 150 -L 100 -l 25 -T 23 -o -p 10`. This corresponds to a window size of 150nt, a maximum base-pair span of 100 nt (as recommended in Chapter 6), a maximum duplex length of 25 nt (no long bulges allowed), a temperature of 23°C, and a seed interaction of at least 10 nt (10 consecutive base-pairs).

[2] Please note that `IntaRNA` [34] sometimes predicts a base to be unpaired at the end of the amiRNA sequence, even though it is complementary to the target site.

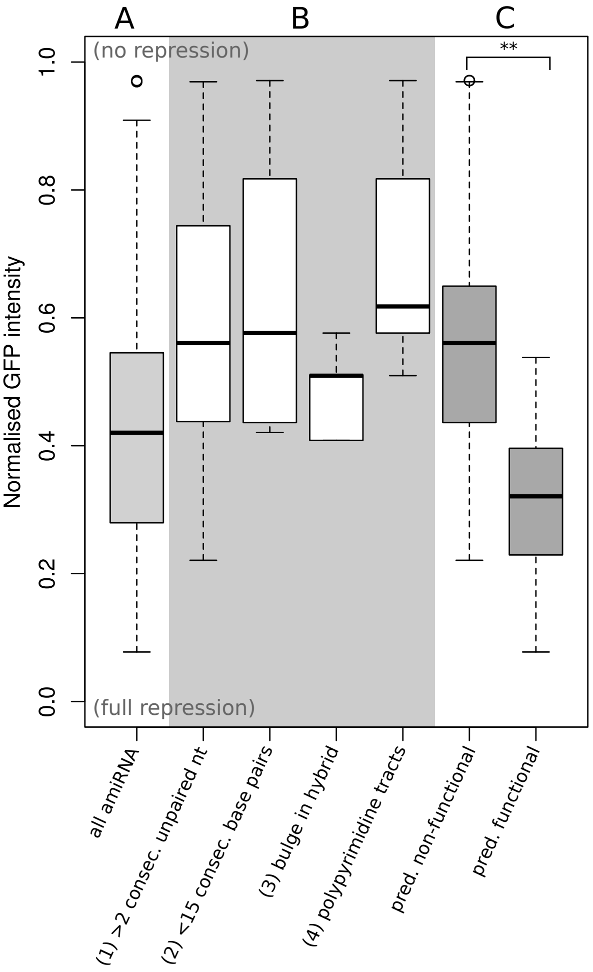


Figure 10.3. Unfavourable amiRNA target–interaction characteristics that led to reduced repression efficacies. (A) Range of all amiRNA efficacies. (B) each boxplot displays amiRNA efficacies only when the indicated unfavourable target-interaction characteristics applies. (C) All amiRNAs for which at least one unfavourable feature applies is predicted to be non-functional, the remaining functional amiRNAs. The distributions of both groups are significantly different (indicated by the two stars) with $p = 6.23 \times 10^{-9}$ (Wilcoxon Rank Sum test) or $p = 7.59 \times 10^{-8}$ (Student's t-test).

We visualised the individual and combined effect of the four binary features on amiRNA efficacies in Figure 10.3. Figure 10.3.A gives the range of efficacies for all amiRNA. Figure 10.3.B summarises the efficacies for amiRNAs for each binary feature individually. Finally, in Figure 10.3.C, we separated the amiRNAs into a set of "predicted" functional and non-functional instances by defining an instance to be non-functional when at least one of the four features in Figure 10.3.B is true. When performing a two-sample statistical test on the normalised GFP intensities of the functional and non-functional sets, they were considered to belong to different distributions with highly significant p-values: $p = 6.23 \times 10^{-9}$ with the Wilcoxon Rank Sum test, or $p = 7.59 \times 10^{-8}$ with the Student's t-test ($t = -6.37$). Results show that a combination of all four features removes most of the amiRNAs with low efficacies. The first three features imply that efficient repression requires sufficient complementarity between the amiRNA and target site to exist; amiRNAs with near-perfect complementarity were most effective. Therefore, the assumption that a perfect complementarity leads to suboptimal repression efficacies is not true for this dataset of amiRNAs. Hence, near-perfect complementarity may lead to more robust results in future amiRNA designs.

The polypyrimidine tracts (fourth feature), targeted by non-functional amiRNA, are potentially motifs for an RNA-binding protein, possibly the polypyrimidine-tract–binding protein (PTB). Since amiRNAs within these regions display the lowest efficacies, it is possible that bound proteins prevent access to the target sites of the respective amiRNA. PTB is a ubiquitous protein that was originally identified as significant for splicing, but also has diverse roles in other cellular processes including polyadenylation, mRNA stability and translation initiation [275]; and homologs exist in *A. thaliana* [270]. Ideally, one would want to search for existing regulatory elements on the mRNA and avoid these regions when designing amiRNAs or any other artificial regulatory mechanism. Although in Part V, we developed an approach for finding such regulatory elements in mRNA and specifically for RBPs [P6], this problem is still far away from being solved computationally. More importantly, large-scale datasets are still lacking in plants. Currently we know of no `CLIP-seq` experiments that have been performed for specific RBPs in plants. Recent `CLIP-seq` protocols designed to map transcriptome-wide binding profiles (in human cells) have been introduced recently [12,38,214] and this information could be integrated into design pipelines. However, also these types of binding data are not yet available for plant cells. Therefore, for subsequent amiRNA design, we filter candidates provided by the WMD tool by the duplex features only. These are easy to compute and no additional data is required.

It is interesting that while accessibility has been shown to sometimes influence miRNA [164] or siRNA [303] gene silencing, in this data, target-site accessibility did not significantly influence repression efficacy. Also not when testing all possible sub-regions of the target site.

10.1.4 Improved amiRNA design

Ren and Dovzhenko optimised the protocol to determine amiRNA efficacy using luminescence instead of fluorescence measurements[1]. In particular, efficacy was measured as the ratio between the luminescence levels of firefly luciferase (measuring repression efficacy) and renilla luciferase (measuring transformation efficiency); the first was translationally fused to the targets as GFP was previously and the second was not influenced by the amiRNA and served as a control. The underlying protoplast system remains the same, however. With this optimised system, ten (coloured green by WMD) amiRNAs were designed using WMD and selected to target the JMJ10 mRNA (TAIR AT1G78280). According to the protoplast detection system, only three out of the ten amiRNAs were considered semi-functional to functional. Hence, we again see that only 30 % of the original amiRNAs are functional. When applying the unfavourable duplex features, determined in Section 10.1.3 (Figure 10.3), to predict amiRNA functionality, nine out of ten predictions were correct; only the amiRNA J10 out of the four amiRNAs that were predicted to have sufficient complementarity did not

[1] See the dissertaion of Fugang Ren for experimental details, titled "*Development of novel technologies for functional characterization and regulation of genes activity in plants*" and submitted in 2014 to the Biology Department of the Albert-Ludwigs-University Freiburg. This work has not been published to date.

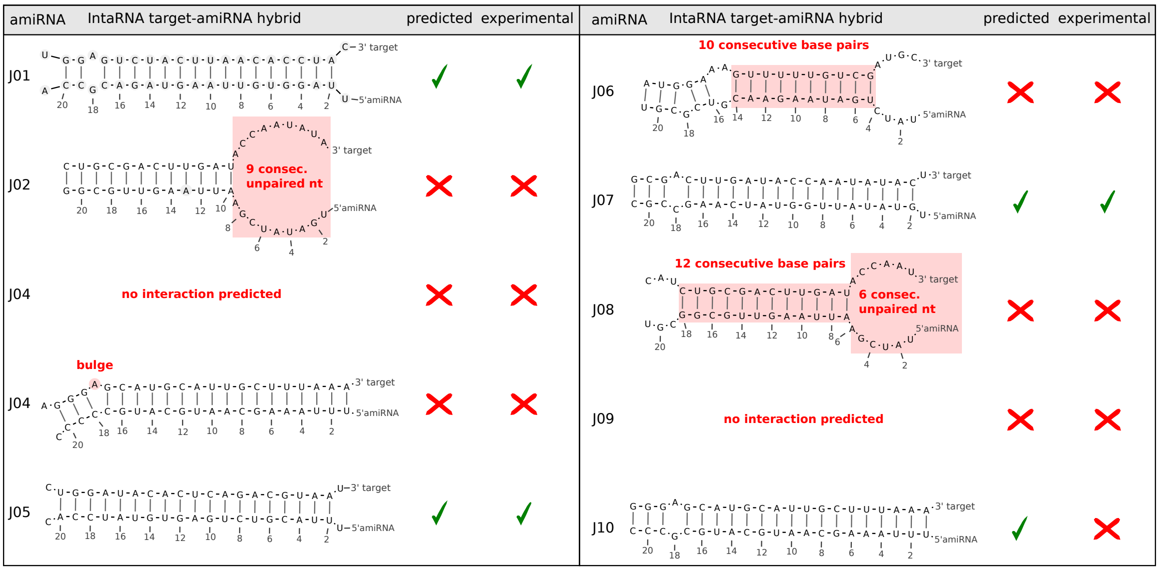


Figure 10.4. Improving design of amiRNA–target interactions. A further ten amiRNAs were designed using WMD against the JMJ10 mRNA. Interactions were predicted using `IntaRNA`[1] and duplex structures were predicted to be non-functional (red cross) if at least one of the three unfavourable features (highlighted in red), determined in Section 10.1.3 , applied and functional (green tick) if it had near-perfect complementarity. Potential regulatory elements on the target mRNA were not considered. The efficacy of the amiRNA was determined by Fugang Ren by luminescence experiments in `A. thaliana` protoplasts. Again only 30 % (3 out of 10) of the WMD-designed amiRNAs were functional, however, with the extended duplex features, we achieved a 90 % prediction accuracy (9 out of 10 correctly predicted).

show any repression activity (see Figure 10.4). Possible reasons for the non-functionality of J10 could be (1) regulatory binding elements situated in the vicinity of the J10 target site or (2) the upstream amiRNA processing or loading into the RISC complex could have failed.

Figure 10.4 also illustrates that `IntaRNA` interaction predictions are likely to differ significantly from `RNAcofold` predictions. For two out of the ten amiRNAs, `IntaRNA` did not predict an interaction that satisfied the energy threshold—even though WMD interactions were designed to have sufficient binding energies that exceed 70 % of their best potential energies. However, `IntaRNA`-predicted duplex structures fitted better to the experimental data. This strongly implies that the more advanced algorithm of `IntaRNA` is more accurate for predicting amiRNA–target interactions.

10.2 The context of artificial binding sites affects repression efficacy

In Chapter 8 (Part V), we determined that there is a significant signal of increased accessibility downstream of MRE sites in *A. thaliana*. Here, we investigated whether the target site context can affect amiRNA repression efficacy as well.

10.2.1 Experimental setup and data processing

From the 62 amiRNA designed to target PIN1 (Section 10.1), we selected one efficient and one inefficient amiRNA, not located near the UC-rich region: P01 with $\mathcal{Q}(P01) = 0.08$ (high repression activity) and P35 with $\mathcal{Q}(P35) = 0.62$ (low repression activity). P01 is fully complementary to its target site, with the exception of the two end nucleotides, P35 has a mismatch at position 19 of the amiRNA, which leads to a predicted unpaired end of three bases (Figure 10.5). To determine the influence of the sequence context of a target site, we inserted the target sites for P01 and P35 into ten different locations in a target mRNA, spread evenly across the coding sequence (CDS); the target site for P35 was changed such that it had full complementarity to P35. The locations for inserting the target site were selected for each target gene as follows:

1. The CDS of the target gene was divided into ten subsequences of equal length.
2. The target sequence for P01 and P35 was inserted into a single position in one subsequence at a random location.
3. Accessibility profiles were calculated using `RNAplfold` [18, 19][1] and plotted as graphs for the CDS with target sequence inserts from both P01 and P35 respectively.
4. Steps (2) and (3) were repeated ten times for each of the ten subsequences of the CDS.
5. A single insertion location for both P01 and P35 target sequences was selected for each subsequence under the condition that the ten selected locations displayed a large variety of accessibility profiles via visual inspection of the profile graphs.

The experiment was performed for four different *A. thaliana* genes that we have not assessed previously: ATGR2 (AT3G54660), NSF (AT4G04910), CDC48B (AT2T03670), and ATDPB (AT5G03415). The experimental protocol for target-site insertions and efficacy measurements can be taken from the dissertation of Fugang Ren[2]. For ATGR2, three independent replicates were produced to compare the reproducibility of measurements derived from the applied protocol (see Figure D.23). Results showed that in general, reproducibility was good for measurements that corresponded to efficient amiRNA repression. Measurements that involved high luminescence ratios, however, did not correlate well. In particular, two replicates for P01 achieved a significant Pearson's correlation coefficient of 0.75 (p=0.01).

The raw luminescence ratios were first normalised to the positive control (PC) and the negative control (NC)[3] using the "relative response ratio" given in Section 10.1.1, Equation 10.1. For PC and NC, we use the average ratio for the control measurements that were on the same

[1] `RNAplfold` from Vienna Package version 1.8.4 and parameters `-noLP -W 100 -L 50 -u 5`.

[2] The dissertaion of Fugang Ren is titled "*Development of novel technologies for functional characterization and regulation of genes activity in plants*" and was submitted in 2014 to the Biology Department of the Albert-Ludwigs-University Freiburg. This work has not been published to date.

[3] The positive control contains an amiRNA that was shown to be active in repressing the firefly luminescence gene, and the negative control contains a vector with the luminescence gene, but no amiRNA to repress it (see dissertation of Fugang Ren for details).

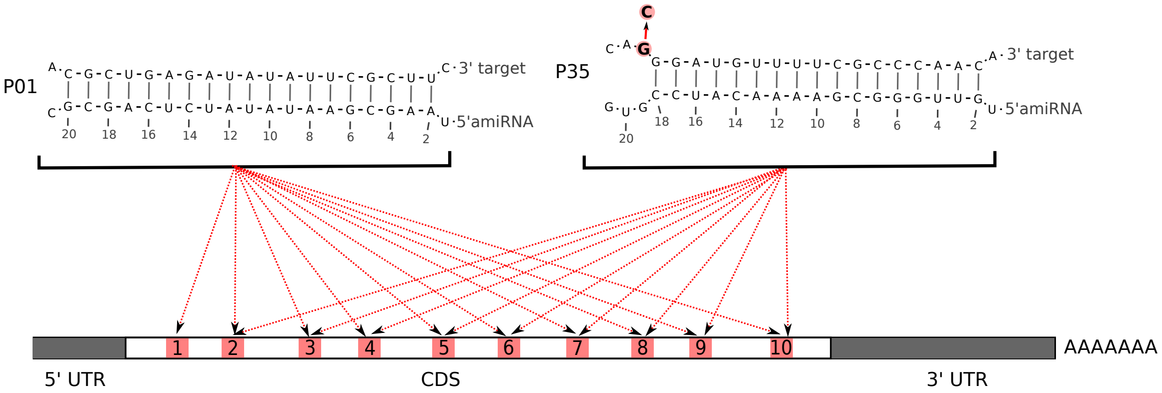

P01
A·C-G-C-U-G-A-G-A-U-A-U-A-U-U-C-G-C-U-U·C·3' target
C·G-C-G-A-C-U-C-U-A-U-A-U-A-A-G-C-G-A-A·U·5'amiRNA
20 18 16 14 12 10 8 6 4 2
P35
C
C·A·G·G-G-A-U-G-U-U-U-U-C-G-C-C-C-A-A-C·A·3' target
G·U·G·C-C-U-A-C-A-A-A-A-G-C-G-G-G-U-U-G·U·5'amiRNA
20 18 16 14 12 10 8 6 4 2
1
2
3
4
5
6
7
8
9
10
AAAAAAA
5' UTR
CDS
3' UTR

Figure 10.5. Controlled experiments for investigating varying target site contexts. The amiRNAs P01 and P35 were selected from the 62 amiRNAs designed to target PIN1 (Section 10.1.3). The depicted duplex structures were predicted using `IntaRNA`. These target sites were extracted as they were predicted for PIN1, except that a single change from G to C at position 19 (relative to the amiRNA) was made so that the target site is fully complementary to P35, which satisfies previously determined duplex features. These target sites were then inserted randomly, but fairly evenly distributed, across a target gene of interest, resulting in ten mRNAs with one inserted target site for each amiRNA. The target positions were selected to reflect a varied accessibility profile as determined by `RNAplfold`. This means that the amiRNA target site remains the same, but the surrounding sequence context is varied. These target site insertions naturally create mutations that considerably change the protein product. Since we are using luminescence reporter genes to measure the amiRNA repression efficacy, these measurements should not be affected significantly by the mutations in the target gene.

plate (in general there were four) and here $I(\alpha)$ is a single luminescence ratio (firefly/renilla). The relative response ratio measures the response in relation to the controls: a value of zero indicates the same repression efficiency as the positive control, and a value of one indicates measurements that are equal to no repression activity. Negative values correspond to efficiencies better than the positive control. This is possible since the positive control is an amiRNA itself and although it gives good results, it is not optimised for functionality, therefore more efficient amiRNAs are possible. Values greater than 1 are due to the large variation in luminescence measurements when no repression activity is measured. For the genes ATGR2 and ATDPB, the negative and positive controls, respectively, are missing, due to unforeseen experimental errors. Since we required not only one, but both, controls for subsequent analyses, we derived the missing control values from all the "valid" pairs of controls. To do this, we calculated the mean ratio of all 24 control pairs: the mean ratio was 0.28 with a variance of 0.02 and a standard deviation of 0.13. The variance of the ratios is very low in comparison with the variance in other raw luminescence measurements, and we thus assume stability of this ratio. According to the mean ratio for ATGR2, we generated values for the negative control by using the original positive control that was on the plates by $NC = PC/0.28$. In this way, we generated four NC values to the four available PC values. Similarly, we computed positive controls for ATDPB by using the negative controls that were available by $PC = NC \times 0.28$. This time six positive control values were computed for the six given negative control values.

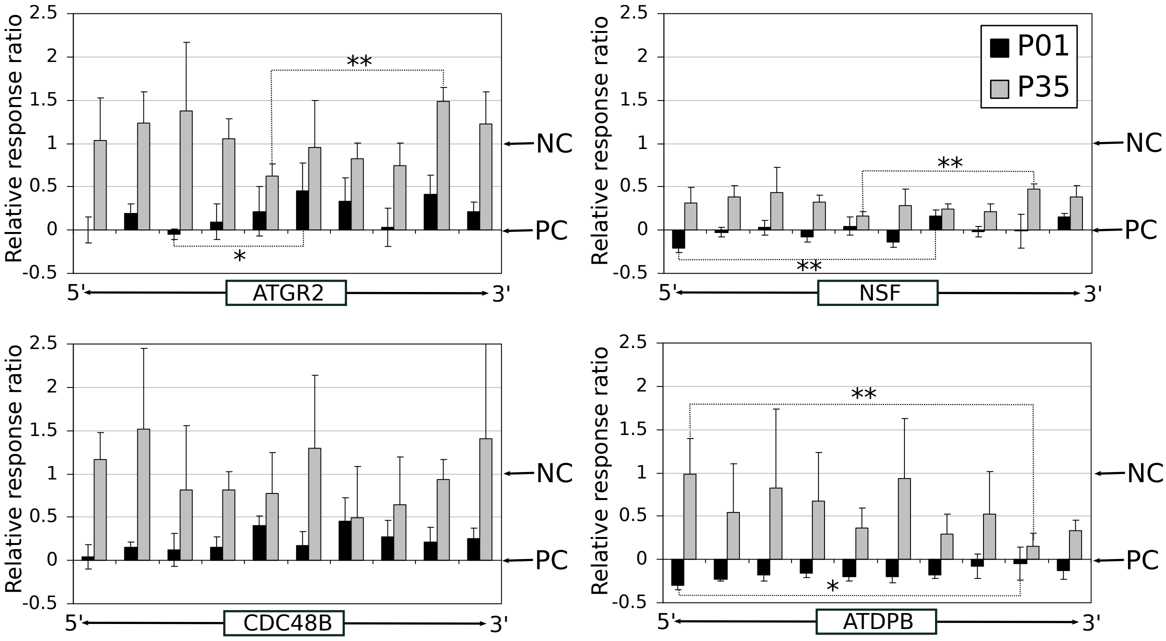


Figure 10.6. The context surrounding target sites significantly affect repression efficacies. The average relative response ratios (y-axis) achieved by P01 and P35 at the ten target site locations, respectively (y-axis); each of the four genes are plotted separately. For the positive and negative controls, the relative response ratio is equal to zero and one, respectively. The error bars correspond to the standard deviation of the replicates. Target-site positions deliver significantly different repression signals, which was tested using a Student's t-test on the replicates between the positions with the maximum and the minimum average signals. A single star denotes a p-value≤ 0.05 and two stars a p-value≤ 0.01.

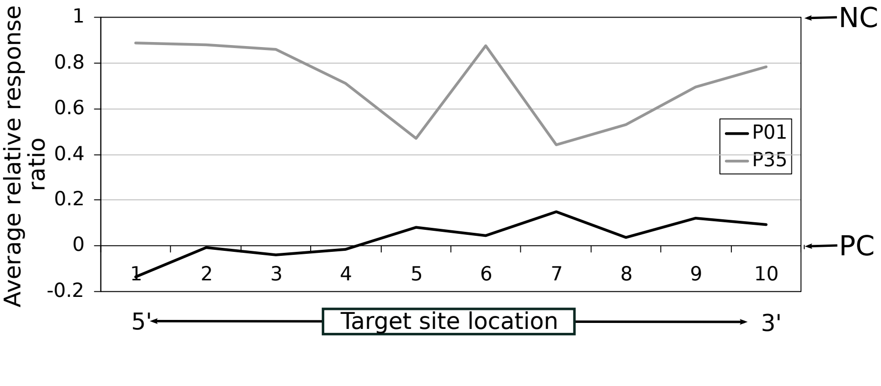


Figure 10.7. Repression efficacies descrease (upward trend) for target sites located towards the 3' end. The average relative response ratio for all replicates of all four genes are plotted separately for P01 (black) and P35 (grey). The functional amiRNA P01 shows an upward trend in relative response ratios (corresponding to a decreasing trend in repression efficacy) when the target-site is located further towards the 3' end of the coding sequence in the respective mRNA. Because of its general inactivity, no such trend is observed for P35.

10.2.2 Results

There are three main results of this experiment: (1) the sequence context of the amiRNA target site affects the repression efficacy to a certain degree, however, (2) despite near-perfect complementarity and various target site locations, P35 is almost always not active or at best only slightly active (Figure 10.6); and (3) for P01, there is a slight but steady decrease in repression efficiencies when target sites are located further towards the 3' end of the coding sequence (Figure 10.7). Since P35 is generally non functional, no such trend is visible.

Repression efficiencies are affected by the context of target site locations for both P01 and P35. Figure 10.6 compares the repression efficiencies of P01 and P35 at the ten target site locations for each of the four genes separately; at each position, the sequence and the structure context surrounding the amiRNA target varies. In general, P01 is still a functional amiRNA and P35 is still not active, however, fluctuations between the individual positions of the target sites in the coding sequence of the mRNA are significant (with p-values ≤ 0.05 in a Student's t-test comparison between positions with minimum and maximum efficacies). For example, in the experiment on the ATGR2 gene, although still functional, P01 shows a significant decrease in repression efficiency at position six. In contrast, the generally inactive P35 displays some repression activity at position five (Figure 10.6). Thus, we can conclude that the sequence and structure context of a target site position has an effect on repression efficiencies for both P01 and P35. Moreover, since P35 was observed to be semi-functional at some positions, its inactivity is not due to incorrect prior processing. Although the target sites of P01 and P35 were inserted at exactly the same locations, we observe no correlation between both their efficiencies for each gene. However, the inserted target-site sequence differs for P01 and P35 and therefore, the RNA structural constraints also change.

10.3 Conclusion

When designing amiRNAs for *A. thaliana* (and maybe other plant species as well), we established that the current standard set by the WMD tool still produces many amiRNAs (up to 70 %) with limited repression activity. The assumption, taken from observation on siRNA, that full complementarity of the amiRNA with its target site does not generate functional amiRNA does not hold on our data. Perfect or close-to-perfect complementarity (with mismatches at both ends) does not negatively affect target repression. We determined that at least 14 consecutive base pairs, not more than two consecutive unpaired bases and no bulges were beneficial to target repression. Furthermore, we identified possible RBP binding sites that sequester target repression almost completely. Therefore, further care must be made to avoid endogenous regulatory elements on target genes when more accurate tools arise in future. Currently, it might be wise to avoid low complexity regions that are rich in only one or two nucleotides—as many RBP binding sites seem to identify such low-complexity regions [P6].

In addition, we provide evidence that variations in the sequence context of a target site can change the efficacy of an amiRNA. Whether these affects are due to RNA structure or endogenous regulatory elements in the vicinity of the target site requires further investigation; no correlation with target site accessibility could be determined. Moreover, designed amiRNA should preferentially target sites located close to the 5' end of the CDS of targeted genes. Further design factors that could affect amiRNA processing or loading into the RISC complex require investigation, since some amiRNA sequences remain non-functional, despite perfect complementarity to the target site and varying sequence contexts.

Part VII

Final remarks

Part VII: Final remarks

There is no real ending. It's just the place where you stop the story.—Frank Herbert

The scientific contribution of this dissertation is separated into five parts, PartII–PartVI. Each part presents data analyses and computational approaches that explore different aspects of post-transcriptional regulatory mechanisms: sequence and structure conservation of ncRNA, expression and processing, local structure prediction and stability, characterisation of regulatory recognition elements, and the design of artificial regulators. In the following, key results and their implications and possible benefit to future work are highlighted, and further work to address current limitations is proposed.

The aim of using conservation in Part II was to group CRISPRs into classes with similar sequence and structure properties and ultimately to identify binding motifs and patterns of associated Cas proteins. The assumption is that identified classes are evolutionarily close. By considering sequence-only and sequence-structure conservation separately, we allowed independent Cas-protein binding motifs to be captured on the CRISPRs. With this sequence-and-structure-conservation information, we comprised an evolutionary map of all available CRISPRs and provided an easy-to-use web server, `CRISPRmap`, which automatically assigns CRISPRs to one of the identified sequence families and/or structure classes, and pinpoints their location in the overall `CRISPRmap` tree. This function is useful to extrapolate information from CRISPRs with known to CRISPRs with unknown functions, and to determine rare CRISPR-Cas systems—as we exemplified in the applications of `CRISPRmap` in Chapter 4. CRISPR-Cas systems are generally classified by looking only at the associated Cas proteins, and CRISPR sequences are not considered when assigning a subtype [123, 201, 202, 324]. However, in initial experiments, we found that CRISPR conservation does not always correlate with CRISPR-Cas subtype annotations. Especially archaeal, and a subset of bacterial, CRISPR-Cas subtypes are linked to frequently closely-related CRISPR sequences, despite being associated with different subtypes. We would like to explore the link between CRISPRs and associated Cas proteins more closely in future. To do this, a key problem to solve would be to accurately and automatically link CRISPRs to their *cas* genes; currently associations are only assigned based on gene locations in the genome, which causes frequent problems. Further future plans involve a better characterisation of type II systems. Type II systems have formed the basis for a new and upcoming technology applied to genome editing [308,328]. However, type II systems are rare (in comparison with the other types) and are thus overpowered in `CRISPRmap` by the vast numbers of type I and III systems.

Extracting expression information from `RNA-seq` data worked very well for crRNAs in Part III. Not only could we establish approximate abundances of crRNAs, processed from a single CRISPR array, but we could also determine exact processing sites, and identify intermediate and mature RNA species. A potential limitation of this approach can lie in the `RNA-seq` protocol used for the transcriptome analysis. First, possible difficulties arise when from the transcriptome, RNAs of only a fixed length are selected for sequencing, which does not capture the exact length of mature crRNAs or intermediate species. Second, poly(A) tails ligated to the 3' ends of transcripts may result in inaccuracy when determining exact 3' ends of RNA species (since naturally occurring terminal *A*s cannot be differentiated from the poly(A) tail). Third, we identified a problem when two unknown nucleotides were ligated to the 5' end of transcripts to remedy sequencing bias, however, these lead to slightly offset 5' end detection and provided problems when detecting exact processing sites. Establishing crRNA expression and processing signals is a prerequisite for biological experiments that investigate aspects of the interference stage of the CRISPR-cas systems. In fact, collaboration partners are currently working on an experiment where we used `RNA-seq` to establish correct processing of an artificial crRNA that includes a substituted spacer designed to target and cleave a region of interest. An `RNA-seq` analysis displayed a processing of the artificial crRNA that was similar to the wild-type crRNAs with the same repeat sequence.

The investigation of many post-transcriptional regulatory processes involves determining local regulatory structure or structural accessibility in long RNAs. Therefore, we put considerable effort into benchmarking available RNA-structure prediction algorithms when applied to long RNAs, and characterising their key parameters (Chapter 6, Part IV). We determined that 100–150 nt represents a reasonable amount of locality to predict local structures accurately. Base pairs that extend beyond this region were often predicted incorrectly and were rare. In addition, we identified a strong bias towards high accessibilities and base-pair probabilities at artificial sequence ends, which should be a particular concern when applying window-based approaches or extracting sequences from a larger context. Predictions at artificial window termini can either be ignored, or setting window sizes that are sufficiently larger than the maximum base-pair span allowed in the local structure prediction approach mitigates biased probabilities. In the second chapter of Part IV, we provided evidence that the surrounding context can hinder the formation of a structured binding motif and therefore prevents its recognition by its *trans* factor, which limits or aborts the expected regulatory function. In particular, these experiments were performed on CRISPR arrays where the repeat formed a small hairpin that was recognised by a Cas6 endoribonuclease and was subsequently cleaved—only if the hairpin motif was predicted to be stable within its sequence context. This result is especially important for experiments that involve artificial constructs comprising of structured regulatory RNA: it is imperative to first assess the stability of the structure in its sequence context. We proposed the use of structure accuracy to measure structure stability.

In Part V, we first presented an empirical analysis of structure accessibility around miRNA-recognition elements (MREs) in *Arabidopsis thaliana.* We identified a region downstream of the MREs that was significantly more accessible in functional vs. non-functional sites. These results were especially interesting because the measured accessibility directly at the MRE sites was not significantly different between the two sets. This implies that MRE-prediction approaches should be evaluating the context for a further factor binding downstream of MRE sites, which requires an accessible region. Later, in Chapter 9, we introduced a machine-learning technique, based on graph kernels to capture miRNA-MRE binding events. Although first prediction performances were promising, further experiments are still required to test how well this framework can solve the extremely difficult task of predicting MREs. However, we know that the framework already works well when applied to RBPs. Further extensions of the framework and thorough benchmarks comparing with state-of-the art approaches for predicting MREs are in progress.

Finally, in Part VI, we explored the application of artificial miRNAs (amiRNAs) to inhibiting the expression of any target gene in plants. Our data shows that amiRNAs require extensive complementarity to their targets to function well. Although this factor alone is not sufficient to ensure good functionality. In a second experiment, we established that the context of the target site can influence the efficiency of an amiRNA to some degree and that best target-site positions are towards the 5' end of the coding sequence. Again, further experiments are required so that we can learn more details about beneficial target-site contexts. In addition, we need to determine what makes an amiRNA non-functional even when the target site is beneficial and it has sufficient complementarity. But to find answers to both these questions, we first require sufficient data to learn from.

In summary, this thesis should provide insightful and detailed knowledge about the individual topics presented—and deliver ideas and approaches for future computational analyses of post-regulatory mechanisms, in particular with regard to RNA-sequence and -structure properties. The quote from Frank Herbert sums up the conclusion of this dissertation succinctly: the presented work has no real ending; it is just time to put it—as it is—to paper.

APPENDIX A

Terms and abbreviations

accessibility	the probability of a nucleotide or a stretch of nucleotides to be unpaired in the structure ensemble of a given sequence	Section 2.5.6
AGO	a member of the Argonaute protein family, which integrates the miRNA into the RISC complex	Section 2.2.1
amiRNA	and artificial microRNA, designed to knock down the expression of its target gene	Chapter 10 (intro)
AUROC	area under the receiver operator characteristic curve	Section 2.3
avg.	average	
base	the differential part of a nucleotide denoted with a single letter: adenosine (A), cytosine (C), guanine (G) and thymine (T) in DNA or uracil (U) in RNA	Section 2.4
base pair	hydrogen bonds forming between bases, typically between G and C, in DNA between A and T and in RNA between A and U; other base pairs are possible, but not as frequent and are generally ignored in secondary structures	Section 2.4
base-pair span	is the distance on the sequence between the two bases of a base pair; the maximum base-pair span allowed in local structure prediction is usually denoted by L	Definition 2.8
bp(s)	base pair(s); the abbreviation is frequently used as a measure of length for DNA segments or for double-stranded RNA	Section 2.4
Cas	CRISPR-associated protein	Section 2.2.2
cas	CRISPR-associated gene	Section 2.2.2
CDS	coding sequence; the coding part of a messenger RNA	Section 2.1.1

***cis*-regulatory element**	elements that are encoded on the *same* molecule that is being regulated, e.g., regulatory structures and binding sites in the UTRs of the mRNA	Section 2.1.1
CLIP-seq	crosslinking immunoprecipitation RNA sequencing; a method to determine the RNA binding sites of a specific RNA-binding protein	Section 2.7.1
CRISPR	clustered regularly interspaced palindromic repeats; involved in prokaryotic defence of genetic material, processed to mature crRNAs	Section 2.2.2
CRISPR array	At a CRISPR locus, the CRISPR array starts with a leader sequence and then contains multiple copies of a repeat sequence, interspaces with variable-length spacer sequences	Section 2.2.2
CRISPR-Cas system	an adaptive prokaryotic immune system that uses a short ncRNA (crRNA) in combination with associated proteins (Cas proteins) to guide the destruction of invading genetic material	Section 2.2.2
crRNA	CRISPR RNA processed into its mature form that guides the destruction of invading genetic material in a prokaryotic immune system	Section 2.2.2
DNA	deoxyribonucleic acid; the molecule that stores genetic information	
expression level	the number of molecules present in the cell at a fixed time point that represent a single gene product (a specific mRNA, ncRNA, or protein)	Section 2.1
FASTA format	the general format used to store DNA, RNA and protein sequences	Section 2.4
Gene expression	is the process by which the functional product of a gene is produced	Section 2.1
GEO	Gene Expression Omnibus; a database that stores mainly expression data of any kind	
GFP	green flourescence protein; frequently used as a reporter or marker gene in *in-vivo* experiments	
IP	immunoprecipitation; a method used in biology for purifying a specific protein by using antibodies	
kb	kilobase, i.e. $1,000$ bases (nucleotides); usually used as a measure of length for RNA or DNA segments	
lncRNA	long non-coding RNA; similar to mRNA in structure, but does not encode proteins	Section 2.1.2
microarray	a solid substrate to which oligonucleotide probes representing usually the complete set of genes of a genome (variations exist) are attached for the detection of gene expression levels	Section 2.1.4
miRISC	an RNA-induced silencing complex (RISC) loaded with a mature microRNA	Section 2.2.1
miRNA	microRNA; a small RNA that generally regulate expression levels of mRNA via various mechanisms of inhibiting translation	Section 2.2.1
MRE	microRNA recognition element; the binding site of a miRNA on an RNA transcript	Section 2.2.1

mRNA	messenger RNA; contains a coding sequence encoding a (or part of a) protein and regulatory tails 5' and 3'UTRs	Section 2.1.1
NC	negative control; no response expected in a biological experiment	
ncRNA	non-coding RNA; generally performs regulatory functions and does not encode a protein	Section 2.1.2
nt	nucleotide (see nucleotide entry); frequently used as a measure for length for an RNA segment	Section 2.4
nucleotide	a single building block of DNA or RNA (see base) that represents one bit of information in the genetic code	Section 2.4
PARS	parallel analysis of RNA structure; a technique to measure the single- and double-strandedness of an RNA called	[165]
PC	positive control; response expected and known in a biological experiment	
post-transcriptional regulation	is the control of gene expression on the level of RNA after transcription and before translation	Section 2.1
protein	made up of amino-acid polymers and performs structural, signal, or regulatory functions	
RBP	RNA binding protein; generally regulate expression levels of mRNA	Section 2.1.3
read	A single sequence that is produced during RNA or DNA sequencing	
RISC	RNA-induced silencing complex; a protein complex involved in miRNA gene-expression regulation	Section 2.2.1
RNA	ribonucleic acid; the functional copy of DNA	
RNA transcript	is an RNA sequence that has been transcribed from its gene segment on the genome	
RNA-seq	A deep sequencing protocol that sequences purified RNA and is thus used to measure transcript abundancies in the gene expression process	Section 2.1.4
RNP	Ribonucleoprotein; a molecule in which RNA and protein are combined to act together	
RRE	regulatory recognition element; the binding site of *trans* factors on a RNA transcript	
siRNA	small-interacting RNA; involved in RNA interference, i.e., in supressing gene expression	Section 2.2.1
structure accuracy	a measure of stability for RNA secondary structure	Section 6.2
transcription	the process by which a segment DNA is copied from the genome and stored as an RNA molecule	
transcriptome	the complete set of RNA transcripts in a cell and the study of the transcriptome is called transcriptomics	Section 2.1.4

***trans* factor**	a *trans*-encoded (not from the same molecule) regulatory element that interacts with the molecule being regulated, e.g., a RBP or an ncRNA that binds to a mRNA to regulate its expression.	
translation	the process by which protein molecules are produced from the information given on the coding region of messenger RNA	
TSS	transcription start site	
UTR	untranslated region; 5'UTR and 3'UTR are at the 5' and 3' terminus of the mRNA, respectively, and do not code for a protein, but frequently contain regulatory signals for *trans*-factor binding	Section 2.1.1
window	usually a subsequence of a fixed length W	Definition 2.17
window-based approach	an approach that performs calculations within only a subsequence, a window of fixed length usually denoted by W; the window slides along the sequence with a certain step size (this step can also be just one nt); results can be summarised or averaged over all windows	
WMD	The Web MicroRNA Designer (`http://wmd3.weigelworld.org`)	[236]

APPENDIX B

Plagiarism declaration

I herewith declare that I have prepared the present work without any unallowed help from third parties and without the use of any aids beyond those given. All data and concepts taken either directly or indirectly from other sources are so indicated along with a notation of the source. In particular I have not made use of any paid assistance from exchange or consulting services (doctoral degree advisors or other persons). No one has received remuneration from me either directly or indirectly for work which is related to the content of the present dissertation.

The work has not been submitted in this country or abroad to any other examination board in this or similar form.

Freiburg, August 22, 2014

APPENDIX C

Detailed statement of contributions

Although I made an effort to reduce the amount of external contribution presented, modern research is always a joint effort; collaborative research was encouraged and required during my Ph.D. years. Therefore, when omitting external contribution would be detrimental to understanding the presented research, it was retained. A point-by-point clarification of personal contributions are made for Part II–VI. Unless otherwise stated, the work presented is my own original research.

All presented work that has been recycled from my own publications was published under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Contributions to Part II

Chapter 3 presents work from [P3]. This work was a joint effort mainly performed by Omer S. Alkhnbashi, Dr. Dominic Rose and myself; we are joint-first co-authors of the original publication and each of us contributed significantly to the scientific content of the work. In particular, the methods for `CRISPRmap` were mainly conceived by myself and D. Rose. O. S. Alkhnbashi contributed more to the methods involved in collecting the CRISPR-Cas data, Dr. Sebastian Will devised an algorithm for constrained clustering to assign single CRISPRs to a sequence family or structure motif, and Dr. Fabrizio Costa gave advice on how to do the first, basic orientation prediction of the CRISPR sequences. All data were acquired and processed by O. S. Alkhnbashi; I thoroughly checked the data for consistency and integrity during its analysis. The `CRISPRmap` software and the web server were implemented mainly by O. S. Alkhnbashi and D. Rose, with most of the technical know-how coming from D. Rose. I helped to design the functionality and layout of the web server, performed the data analyses (with some support from O. S. Alkhnbashi), generated all figures in Chapter 3, and performed the required literature research. Figures D.1 to D.8 were generated by O. S. Alkhnbashi under my guidance. Moreover, I planned and wrote the article with minor contributions from Prof. Dr. Rolf Backofen, D. Rose and S. Will: any text passages extracted from the original manuscript [P3] and included in this thesis were written by myself. I hereby acknowledge the significant contribution from O. S. Alkhnbashi to `CRISPRmap`, and that it will likely also be a part of his dissertation.

The Chapter 4 on applications and limitations of `CRISPRmap` was devised solely by myself and I also performed all the work that was involved—aside from developing the software and designing the improved orientation prediction of CRISPR sequences in Section 4.3.2, and published in [P1]. My contribution to [P1] was of a minor nature: I helped to design and conceive the research and edited the final manuscript. O. S. Alkhnbashi performed all software development and analyses and created version 2.0 of `CRISPRmap`. F. Costa supervised the project and gave input into the machine-learning and parameter optimisation aspects of the work. The figure included in Chapter 4 was generated by myself. Currently, the web server for `CRISPRmap` is maintained by O. S. Alkhnbashi and Dr. Martin Mann.

Contributions to Part III

All work and figures generated for Part III (Chapter 5) is my own work, based on the publication [P10]. All text passages taken from this publication were written by myself. All wet-lab experiments were devised and performed by Prof. Dr. Wolfgang R. Hess and members of his lab (in addition to creating Figure D.13), in particular, Dr. Ingeborg Scholz performed most of the experiments and Stefanie Hein completed experiments when I. Scholz was on maternity leave. Prof. Dr. R. Backofen devised the idea of finding correlations in the `RNA-seq` data with structural attributes, developed some of the initial software, and oversaw the project.

Contributions to Part IV

Chapter 6 presents the work from [P4]. In this publication, I share the first-author position—equal scientific contribution—with Daniel Maticzka. Although most ideas arose from joint discussions and D. Maticzka and I performed the majority of the work: I put more emphasis into producing the results for Figures 6.1, 6.2, 6.5, 6.6 and 6.7 and D. Maticzka put more emphasis into producing the results for Figures 6.3, 6.4 and 6.8. In specific terms, my work involved the generation and extended processing of the benchmarking dataset of structured *cis*-regulatory elements and performing comparisons on this data. D. Maticzka focussed more on displaying the detrimental effect of artificial border termini, implementing the modification to `RNAplfold` that rectifies this problem, `LocalFold`, and performing comparisons using accessibility data. We both wrote the bulk of the manuscript, some of which has been copied into this dissertation, thus possible overlap with his dissertation is duly noted. Dr. Mathias Möhl provided expertise on partition-function algorithms and checked the validity of presented equations. Joshua Gagnon implemented the web site for the `CisReg` data and Dr. Chris Brown provided additional expertise on *cis*-regulatory elements in mRNAs and provided the list of viable elements to use. Prof. Dr. R. Backofen supervised the project, and formulated the theory of structure accuracy for local structure prediction in Section 6.2.

The work presented in Chapter 7 is based on a collection of published articles [P5,P7,P8,P10, P11]; I am a joint first author of [P10]. For the experimental parts of this chapter, we worked in close collaboration with members of the labs of Prof. Dr. W. R. Hess, Prof. Dr. Anita Marchfelder, and Prof. Dr. Ruth A. Schmitz-Streit. I did not perform any biological experiments, therefore, experimental details are not within the scope of this thesis and should be taken from the original publications. All methods, figures and text were produced by myself (except for Figures 7.3B and D.20, which were created by O. S. Alkhnbashi under my guidance). I developed all software required for this work. Prof. Dr. R. Backofen supervised the work and aided in the representation of Figure 7.5.

Contributions to Part V

The analysis and all methods in Chapter 8 were devised solely by myself. I also performed most of the data analysis with two minor exceptions: the miRNA interaction data from *A. thaliana* was generated by my student assistant, Peter Zeller. The extended analyses on human and firefly miRNA and artificial siRNA in Section D.4 and presented in Figure D.21 was performed by Kyanoush S. Yahosseini during his Bachelor thesis. Both students were supervised closely by myself; thus, all methods and approaches were developed in discussions with me.

Work presented in Chapter 9 is the result of many collaborative participants. First the basic encoding of structured RNA and in particular RREs was developed in equal measures with Dr. Steffen Heyne, D. Maticzka, Dr. F. Costa and myself. In particular, the script for converting `RNAshapes` output to graphs was developed by myself and S. Heyne. D. Maticzka extended it to the viewpoint notion, and I implemented the hypergraph encoding to include the abstract secondary structure element annotation. F. Costa is the author of the NSPD Kernel and has developed the `EDeN` software package to perform feature extraction, machine learning and performance analyses. In addition, F. Costa was the driving force behind the idea of applying the NSPD Kernel to RNA and guided all steps of model and encoding design. The extension of the graph encoding to miRNA-MRE interactions, the modelling of these interactions, and the processing of all miRNA-related data was performed by Michael Uhl during his team project, Master thesis, and as a student assistant, and was always supervised by myself and F. Costa. The basic RNA encoding without viewpoint or hypergraph encodings was published by S. Heyne and F. Costa in [136] and its later modification and application to RBPs, which does not include any of the miRNA-specific encodings, was published in [P6].

For [P6], I was a major contributor to writing and editing the manuscript. The extension to miRNAs is currently unpublished.

Contributions to Part VI

All computational data analyses and all figures presented in Part VI, Chapter 10 were performed and generated by myself and supervised by Prof. Dr. R. Backofen. A minor contribution was given by Dr. Anke Busch who generated the thousands of features for the amiRNA dataset in Section 10.1. The final results presented here were based on the the knowledge I gained from analysing her calculated data. In addition, D. Maticzka provided some aid in originally looking at the fluorescence data and plotting RNA structure features of target sites. In the same way, his results are not directly presented, but his input helped to develop the final work. The experimental aspects of this work were devised and performed by Dr. Claude Becker, Fugang Ren, and Dr. Alexander Dovzhenko under the supervision of Prof. Dr. Klaus Palme. All work in this part is currently unpublished.

APPENDIX D

Supplementary material

D.1 Part II

D.1.1 Additional methods for `CRISPRmap`

This material was taken from the supplement of [P3] and was included here for comprehensive purposes.

Cas subtype annotation from Haft *et al.* 2005.

To annotate the early Cas subtypes from Haft *et al.* [123], we followed the procedure given in Kunin *et al.* [180]. More specifically, we downloaded the single *cas* gene models created by Haft *et al.* from the `TIGRFAM` database. Using the `HMMER` program with the `TIGRFAM` models (same as for the single *cas* gene annotation), we searched the 20 kb of nucleotides up- and downstream of the array locus and annotated a *cas* gene if it was found with an E-value ≤ 0.001. We used a strict annotation of Cas subtypes, whereby all *cas* genes of a subtype were required.

Web server input: adding new repeat sequences to the existing CRISPR clustering

The user of our `CRISPRmap` web server can enter any CRISPR sequences and they will be assigned to our sequence families and structure motifs, if possible, and integrated into the hierarchical `CRISPRmap` tree. Thus, information on conservation is available for not only sequences in our dataset, but also novel, yet unsequenced, CRISPRs. In the following, we describe the procedure for one input sequence, many sequences are done simultaneously in the same way:

1. *Is the repeat sequence in our database?* If the given repeat sequence is in our database, in either orientation, we highlight this sequence (or one if many copies exist) in our `CRISPRmap` cluster tree, and automatically assign it to the corresponding structure motif and/or sequence family and stop here.

2. *What is the correct orientation?* If the user is not sure about the correct repeat orientation, i.e., the checkbox for repeat orientation has been activated, we first predict the orientation with our model described in the methods section of the main manuscript. The orientation should then be consistent with our data.

3. *Is it structured or unstructured?* The RNA structure prediction algorithm, `RNAfold` [140] is used to determine whether the repeat sequence is structured or unstructured. If the minimum free energy structure is the unstructured sequence, i.e., contains no base-pairs, it remains unassigned to a structure motif and we continue with Step 5.

4. *Does it belong to a structure motif?* Albeit a structure being predicted, the repeat does not necessarily belong to a conserved structure motif. We add the repeat sequence to all repeats assigned to one of our structure motifs and re-run `RNAclust` [339] with a modified `UPGMA` algorithm (see following section “Constrained Clustering”). In short, the modification allows the generation of the cluster tree by keeping the motifs intact, i.e. non-overlapping. If a repeat falls into or next to one of the existing structure motifs, we assign it to the motif by the following: (1) The repeat is folded by `RNAfold` [140] with the option -p to calculate a structure dotplot. (2) This dotplot is aligned with the consensus dotplot of the structure motif using `LocARNA`. (3) The repeat is assigned to be a member of the motif if it is able to fold into the consensus structure of that respective motif with at most one base-pair missing. We ensure that the new consensus structure contains at least four base-pairs and is at the same position as previously. A comparison of the new and old consensus structures and alignments is given on the web server results page.

5. *Does it belong to one of our conserved sequence families?* We assign the repeat to a conserved sequence family by comparing it to the previously calculated `ClustalW` sequence profiles [311], see Methods section “Clustering of repeat sequences into conserved sequence families”. Let $sim(F, r)$ be the profile score of a repeat r compared with the profile of the family F, where $r \notin F$. For each family, the minimum F_{min} and maximum F_{max} profile similarity was determined by removing each sequence from the family, re-calculating the profile for the remaining sequences, and determining the similarity score of the respective repeat to the profile. A repeat r was then assigned to a sequence family F if (1) $sim(F, r)$ is greater or equal to F_{min} and (2) the distance between $sim(F, r)$ and F_{max} is the minimum for all families.

6. *Where is it located in the `CRISPRmap` cluster tree?* With a final run of `RNAclust` on all repeat sequences, we get the updated `CRISPRmap` cluster tree and we highlight the input sequence location in this tree. Any additional annotations (outer rings), such as Cas subtype, are not displayed for novel repeat sequences.

Constrained Clustering

We consider the general problem to cluster a set of taxa hierarchically based on their distances. Additionally, we constrain the clustering such that certain, e.g. a priori known, clusters are prevented from mixing with each other.

Given is a set of taxa, indexed from 1 to n, together with all pairwise distances between the taxa; furthermore, a set $\mathcal{X}$ of disjoint clusters of these taxa, i.e., $\mathcal{X}$ is contained in the powerset of $\{1, \ldots, n\}$ and all non-identical clusters c and d in $\mathcal{X}$ do not intersect. Commonly, $\mathcal{X}$ covers only a subset of all taxa; therefore, we distinguish *constrained taxa* (that are contained in some element of $\mathcal{X}$) and the remaining *unconstrained taxa*.

We aim to construct a cluster tree of the taxa, i.e., a rooted binary tree T with n leaves corresponding to the n taxa. First, this tree should reflect the given distances. Second it has to support the clustering given by $\mathcal{X}$ such that clusters in $\mathcal{X}$ are grouped together but unconstrained taxa can be interspersed freely. For this purpose, we require that no subtree of T contains leaves from two different clusters in $\mathcal{X}$ unless both clusters are completely contained in the subtree. We call this condition *$\mathcal{X}$-cluster constraint.* (Formally: for each subtree with leaves L and each pair of non-identical clusters c and d in $\mathcal{X}$, $c \cap L \subset c$ implies $d \cap L = \emptyset$.)

Our novel constrained clustering algorithm is based on the unweighted pair group method `UPGMA`. The original algorithm `UPGMA` starts from n singleton clusters corresponding to the n taxa. Until all clusters are combined, it iteratively merges the two nearest clusters. For the latter, the cluster distances are initially derived from the input distances and distances to new clusters are computed after each merge of clusters. The sequence of merges determines the cluster tree. The novel algorithm modifies `UPGMA`, such that, in each iteration, it merges the nearest pair of clusters that can be merged without violating the $\mathcal{X}$-cluster constraint. To check this condition efficiently, we keep track for each cluster whether it contains some elements of a cluster in $\mathcal{X}$ and whether it includes such a cluster completely. Merging two clusters does violate the constraint if and only if each cluster overlaps some cluster in $\mathcal{X}$ but does not cover it completely.

Horizontal gene transfer between bacteria and archaea

Although archaeal CRISPRs are generally well-separated from bacterial ones in general, we observed a few instances where an archaeal CRISPR is located within a bacterial-dominated region and vice versa. To investigate whether these mixed regions could arise from potential horizontal transfer, we applied `BLAST` [5] to search for homologous Cas1 (or Cas2) protein sequences (Cas1 and Cas2 are the most ubiquitous Cas proteins and exist in both bacteria and archaea). We identified 24 archaeal and 8 bacterial repeats that were assigned to sequence families or structure motifs dominated by the opposite domain. For 75 % (18 out of 24) of the archaeal repeats, we identified Cas1 or Cas2 homologs in bacteria in the top five `BLAST` hits (E-value $\leq 2 \times 10^{-10}$); the same was true for only one of the four bacterial repeats.

D.1.2 Supplementary tables for CRISPRmap

Number of Cas subtype annotations

We annotated each CRISPR in our dataset according to the closest Cas subtypes as described in the methods of the manuscript. The two major Cas subtype annotation systems were considered [123, 202]; the number of CRISPRs we annotated with each subtype is given in Table D.1.

Subtype	Archaea	Bacteria	Total
10 subtypes from Makarova *et al.* 2011 [202]			
I-A	134	203	337
I-B	89	293	382
I-C	**14**	**322**	**336**
I-D	49	38	87
I-E	**8**	**447**	**455**
I-F	**1**	**155**	**156**
II-A	0	50	50
II-B	9	95	104
III-A	148	223	371
III-B	108	149	257
% CRISPR	87 %	68 %	72 %
8 subtypes from Haft *et al.* 2005 [123]			
Apern	65	0	65
Dvulg	**1**	**184**	**185**
Ecoli	**8**	**369**	**377**
Hmari	15	36	51
Mtube	8	9	17
Nmeni	0	27	27
Tneap	89	254	343
Ypest	**0**	**120**	**120**
% CRISPR	29 %	35 %	34 %

Table D.1. The number of identified Cas subtype annotations for our REPEATS dataset. There were double as many annotations using the more recent classification from Makarova *et al.*, however, we did not require that all *cas* genes from the respective subtype to be present; whereas the annotations performed for Haft *et al.* were more strict, since we used full subtype models (see methods). In general, Dvulg, Ecoli, Hmari, Mtube, Nmeni, and Ypest correspond to I-C, I-E, I-B, III-A, both type II, and I-F, respectively. Structured repeats with very stable and conserved hairpin motifs, mainly found in bacteria, are written in bold. Note that the 9 subtype II-B CRISPRs in archaea are likely to be incorrect as we did not identify an RNase III in these organisms. Automated annotation of subtype II-B was especially difficult as it contains no subtype-specific Cas protein.

Summary tables of sequence families and structure motifs

Supplementary Tables D.2–D.19 summarise the sequence families and structure motifs, sorted according to the superclass they belong to. The numbering of the families is according to the number of repeats belonging to that family. The annotations in each column is done manually with respect to the majority of repeats in that family (see other supplementary file for the full list). For the Cas subtype, an annotation is only given if this is more or less clear. If there is a complete mix of subtypes, no information is given. The Cas subtypes are summarised according to the *cas* genes that are found in the majority of chromosomes which contain the CRISPRs of each family or motif. More details of the majority *cas* genes is given on the web server. Archaeal families and motifs are highlighted in blue. If the `CRISPRmap` web server is updated in future, then these tables supply a record for sequence families and structure motifs that are referred to in this work. The secondary structures of the motifs and sequence logos of the families are also provides in the tables.

Table D.2. Summary for the bacterial sequence families in Superclass A.

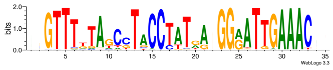
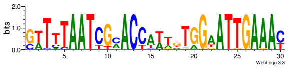
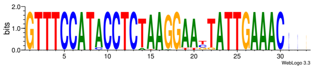
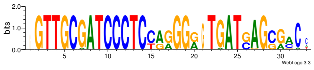
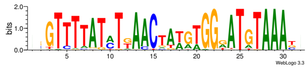
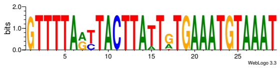
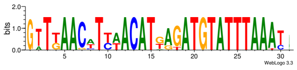
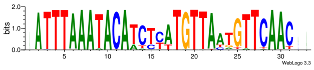
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F1		289	M10 unstructured	Firmicutes	I-B III-A III-B
F25		23	unstructured	mixed bacteria	I-A II-B III-A
F16		40	unstructured	Thermotogae	III-A
F30		19	M2	Actinobacteria	-
F6		124	M8 unstructured	Firmicutes	I-A
F28		20	unstructured	Firmicutes	I-A
F34		15	M21	Firmicutes	II-B
F9		76	M7	Firmicutes	III-B

Table D.3. Structure motif summary for bacterial motifs in Superclass A.

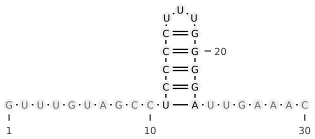
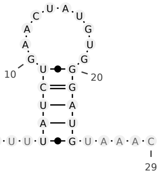
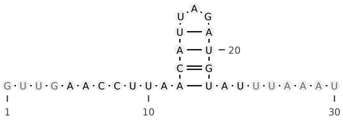
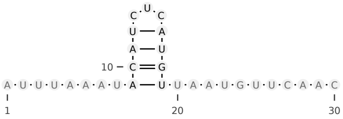
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M10	 U U U C G C G 20 C G C G G U U U G U A G C C U A U U G A A A C 1 10 30	50	F1	Firmicutes	I-B II-B III-A
M8	 C U A U A G A U G G 10 U G 20 C G U A A U G U U U U G U A A A C 1 29	55	F6	Firmicutes	I-A I-B III-A
M21	 U A G U A A U 20 C G G U U G A A C C U U A A U A U U U A A A U 1 10 30	26	F34 unassigned	Firmicutes	-
M7	 C U C U A A U 10 C G A U U U A A A U A U U A A U G U U C A A C 1 20 30	78	F9	Firmicutes	I-A III-B

Table D.4. Summary for the archaeal sequence families in Superclass A.

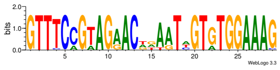
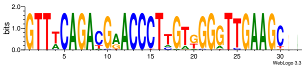
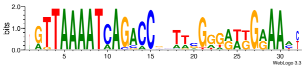
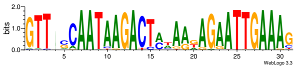
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F29		20	un-structured	Euryarchaeota Crenarchaeota	III-A
F19		32	un-structured	Euryarchaeota	-
F7		108	M15 M16 M27	Euryarchaeota	I-A
F10		70	un-structured	Euryarchaeota	I-B

Table D.5. Structure motif summary for archaeal motifs in Superclass A.

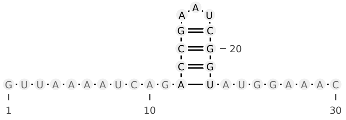
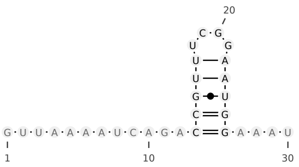
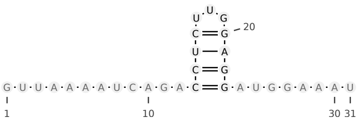
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M15		35	F7	Euryarchaeota	-
M27		17	F7	Euryarchaeota	-
M16		33	F7	Euryarchaeota	-

Table D.6. Sequence family summary for Superclass B.

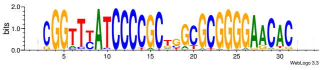
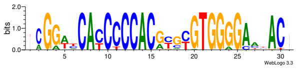
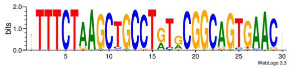
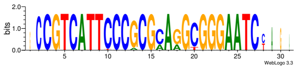
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F2		221	M1	Actinobacteria Proteobacteria	I-E
F18		35	M1	mixed bacteria	I-E II-B
F8		88	M6	Proteobacteria	I-F
F22		26	M18	Proteobacteria	I-E

Table D.7. Structure motif summary Superclass B.

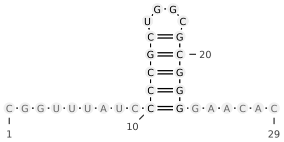
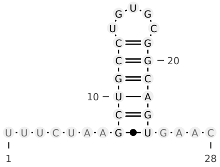
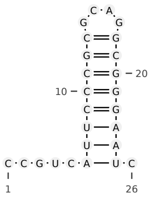
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M1		265	F2 F18	mixed bacteria	I-E
M6		89	F8	Proteobacteria	I-F
M18		28	F22	Proteobacteria	III-B

Table D.8. Sequence family summary for Superclass C.

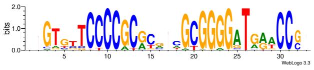
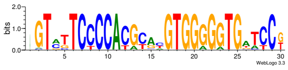
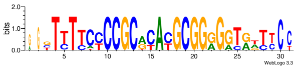
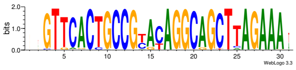
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F4		172	M2	Actinobacteria Proteobacteria	I-C I-E II-B
F21		27	M2	mixed bacteria	I-E
F33		16	M2	mixed bacteria	I-C I-E II-B
F5		135	M4	Proteobacteria	I-F

Table D.9. Structure motif summary for Superclass C.

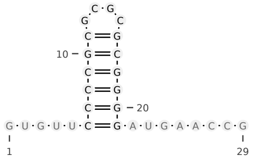
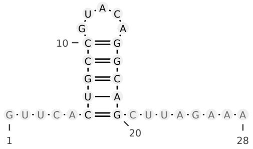
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M2	 C·G G C C=G 10 – G=C C=G C=G C=G – 20 G·U·G·U·U·C=G·A·U·G·A·A·C·C·G 1 29	222	F4 F21 F30 F33 unassigned	mixed bacteria	I-E
M4	 U·A·C G A 10 – C=G C=G G=C U—A G·U·U·C·A·C=G·C·U·U·A·G·A·A·A 1 20 28	142	F5 unassigned	Proteobacteria	I-F

Table D.10. Sequence family summary for Superclass D.

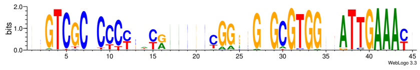
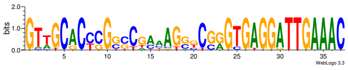
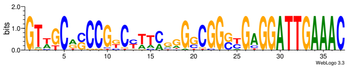
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F3		210	M3 M9	mixed bacteria	I-C
F37		14	M9	Deinococcus-Thermus	I-C III-B
F32		18	M9	Deinococcus-Thermus Proteobacteria	I-C

Table D.11. Summary for structure motifs in Superclass D with sequence conservation.

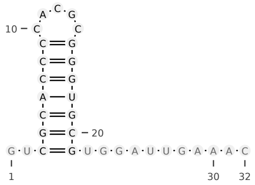
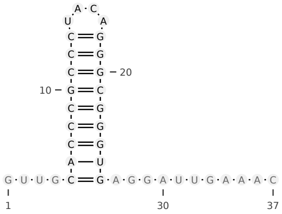
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M3		195	F3	mixed bacteria	I-C
M9		52	F3 F32 F37	mixed bacteria	I-C I-A

Table D.12. Summary for structure motifs in Superclass D without sequence conservation.

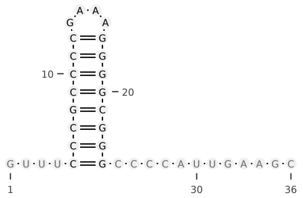
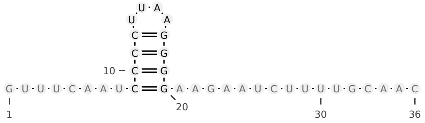
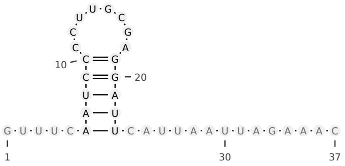
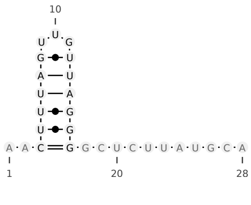
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M19	 A·A G A C=G C=G 10 – C=G C=G – 20 G=C C=G C=G G·U·U·U·C=G·C·C·C·C·A·U·U·G·A·A·G·C 1 30 36	28	unassigned	mixed bacteria	I-A II-B III-B
M25	 U·A U A C=G C=G 10 – C=G G·U·U·U·C·A·A·U·C=G·A·A·G·A·A·U·C·U·U·U·U·G·C·A·A·C 1 20 30 36	19	unassigned	mixed bacteria	III-A III-B
M30	 U·G U C C G C A 10 – C=G C=G – 20 U—A A—U G·U·U·U·C·A—U·C·A·U·U·A·A·U·U·A·G·A·A·A·C 1 30 37	13	unassigned	Cyanobacteria Chloroflexi	I-E II-B
M33	 10 U U G G●U A—U U—A U●G U●G A·A·C=G·G·C·U·C·U·U·A·U·G·C·A 1 20 28	10	unassigned	mixed bacteria	II-B

Table D.13. Sequence family summary for Superclass E.

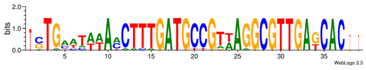
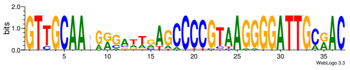
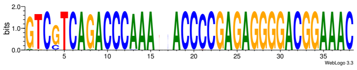
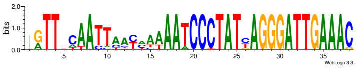
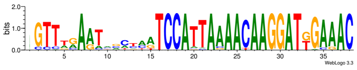
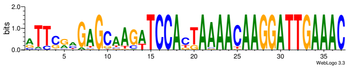
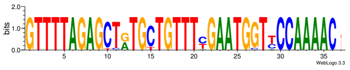
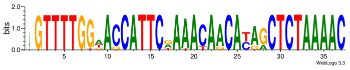
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F39		13	M5	mixed bacteria	I-A I-B II-B
F31		19	M5	Deinococcus-Thermus	III-A
F12		45	M5	Actinobacteria	II-B III-A
F23		24	M12	Cyanobacteria	I-D II-B
F20		28	M13 unstructured	Euryarchaeota mixed bacteria	I-B
F26		23	M13 unstructured	Euryarchaeota	-
F35		15	unstructured	Firmicutes	II-A
F27		22	M14 unstructured	Firmicutes	II-A II-B

Table D.14. Summary of bacterial structure motifs in Superclass E with sequence conservation.

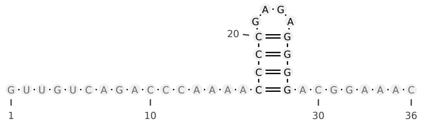
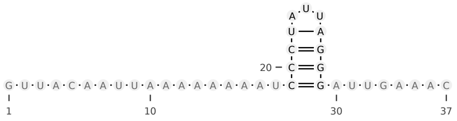
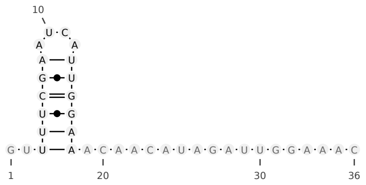
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M5	 A·G G A 20 C=G C=G C=G G·U·U·G·U·C·A·G·A·C·C·C·A·A·A·A·C=G·A·C·G·G·A·A·A·C 1 10 30 36	106	F12 F31 F39 unassigned	Cyanobacteria mixed bacteria	II-B III-A
M12	 A U U U—A C=G 20 C=G G·U·U·A·C·A·A·U·U·A·A·A·A·A·A·A·A·U·C=G·A·U·U·G·A·A·A·C 1 10 30 37	40	F23 unassigned	mixed bacteria	-
M14	 10 U·C A A A—U G•U C=G U•G U—A G·U·U—A·A·C·A·A·C·A·U·A·G·A·U·U·G·G·A·A·A·C 1 20 30 36	35	F27 unassigned	Firmicutes Cyanobacteria	II-A II-B

Table D.15. Summary of bacterial structure motifs in Superclass E without sequence conservation.

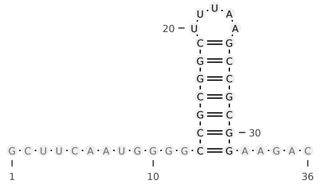
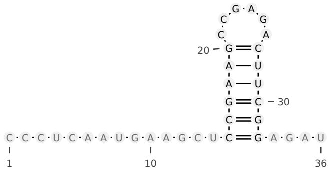
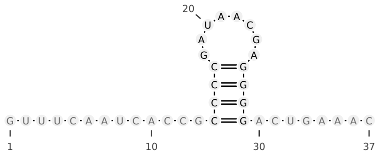
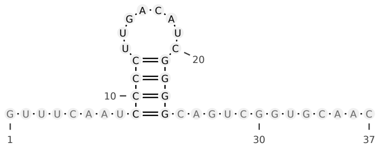
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M23	 U·U·A 20 – U A C=G G=C G=C C=G G=C C=G – 30 G·C·U·U·C·A·A·U·G·G·G·G·C=G·A·A·G·A·C 1 10 36	23	unassigned	mixed bacteria	-
M26	 G·A C G C A 20 – G=C A—U A—U G=C – 30 C=G C·C·C·U·C·A·A·U·G·A·A·G·C·U·C=G·A·G·A·U 1 10 36	19	unassigned	Actinobacteria	-
M28	 20 U·A·A·C A G G A C=G C=G C=G G·U·U·U·C·A·A·U·C·A·C·C·G·C=G·A·C·U·G·A·A·A·C 1 10 30 37	16	unassigned	mixed bacteria	I-C III-A
M24	 G·A·C·A U U U C 20 C=G C=G 10 – C=G G·U·U·U·C·A·A·U·C=G·C·A·G·U·C·G·G·U·G·C·A·A·C 1 30 37	21	unassigned	mixed bacteria	-

Table D.16. **Summary of archaeal structure motifs in Superclass E.**

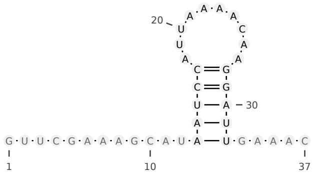
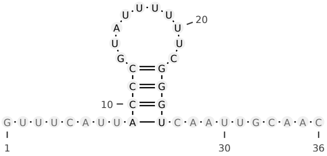
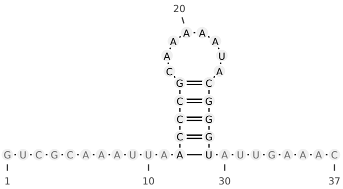
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M13		37	F20 F26	Euryarchaeota	I-A
M31		11	unassigned	Euryarchaeota mixed bacteria	-
M29		14	unassigned	Euryarchaeota mixed bacteria	II-B

Table D.17. Sequence family summary for Superclass F.

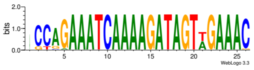
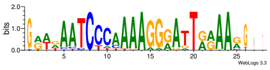
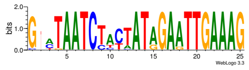
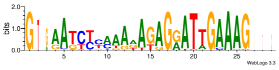
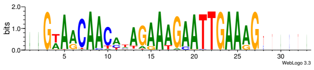
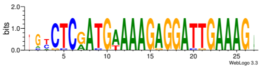
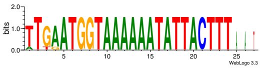
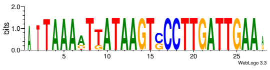
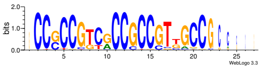
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F24		23	unstructured	Crenarchaeota	III-A III-B
F15		42	M22 unstructured	Crenarchaeota	I-A III-B
F13		44	M17 unstructured	Crenarchaeota	I-A III-B
F11		49	M11 unstructured	Crenarchaeota	III-B
F14		44	unstructured	Crenarchaeota	I-A I-D III-A
F38		13	unstructured	mixed archaea	I-A III-B
F36		15	M20	Firmicutes	-
F40		13	unstructured	Proteobacteria	I-B
F17		39	unstructured	Actinobacteria	-

Table D.18. Summary for archaeal structure motifs in Superclass F.

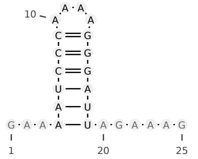
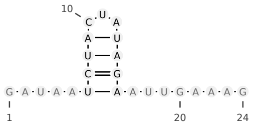
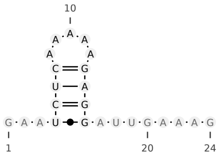
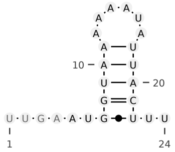
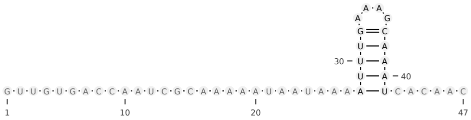
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M22	 A·A 10 A A C═G C═G C═G U—A A—U G·A·A·A—U·A·G·A·A·A·G 1 20 25	24	F15	Crenarchaeota	I-A III-B
M17	 10 C·U·A A—U U—A C═G G·A·U·A·A·U—A·A·U·U·G·A·A·A·G 1 20 24	29	F13	Crenarchaeota	I-A III-B
M11	 10 A·A·A A A C═G U—A C═G G·A·A·U●G·A·U·U·G·A·A·A·G 1 20 24	45	F11 unassigned	Crenarchaeota	III-A III-B
M20	 A·A A U A A A—U 10 A—U U—A 20 G═C U·U·G·A·A·U·G●U·U·U 1 24	27	F36 unassigned	Firmicutes Crenarchaeota	-

Table D.19. Final structure motif unassigned to a Superclass.

#	Structure Motif	Size	Families	Taxonomy	Subtypes
M32	 A·A A G G═C U—A 30 U—A U—A 40 G·U·U·G·U·G·A·C·C·A·A·U·C·G·C·A·A·A·A·A·U·A·A·U·A·A·A·A—U·C·A·C·A·A·C 1 10 20 47	10	unassigned	Becteroidetes	II-B

D.1.3 Supplementary figures for `CRISPRmap`

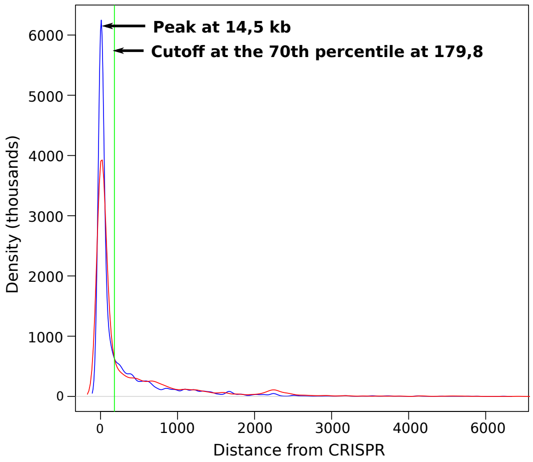


Figure D.1. Distance of *cas* genes in the annotation of subtypes from Makarova *et al.* 2011. Distance of signature subtypes is in blue and the distance of signature types is in red; the cutoff is indicated with the green line. The plot shows the distribution of the closest signature genes to the CRISPR array. A signature gene is one that is unique to either the subtype or the type, respectively. Figure taken from [P3].

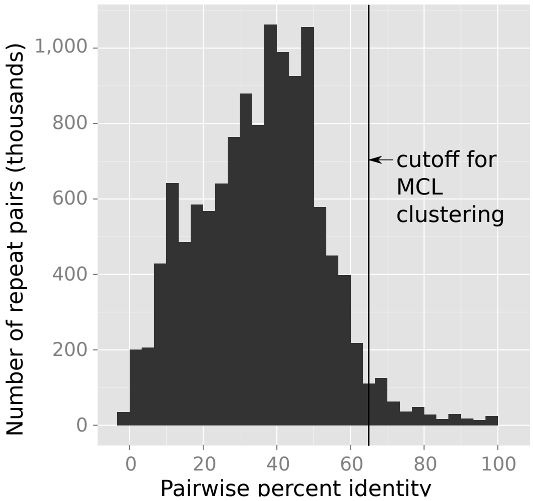


Figure D.2. Pairwise similarities for repeats. We plotted the distribution of pairwise percent identities (x-axis) of Needleman-Wunsch [230] alignments for all repeats to determine a cutoff for the Markov clustering (MCL). Here we see that 65 % is a reasonable cutoff in comparison with the background distribution. Repeats with a similarity below 65 % are set to zero. Because of the short repeat length and conserved sequence motifs, it is necessary to choose such a high cutoff. Figure taken from [P3].

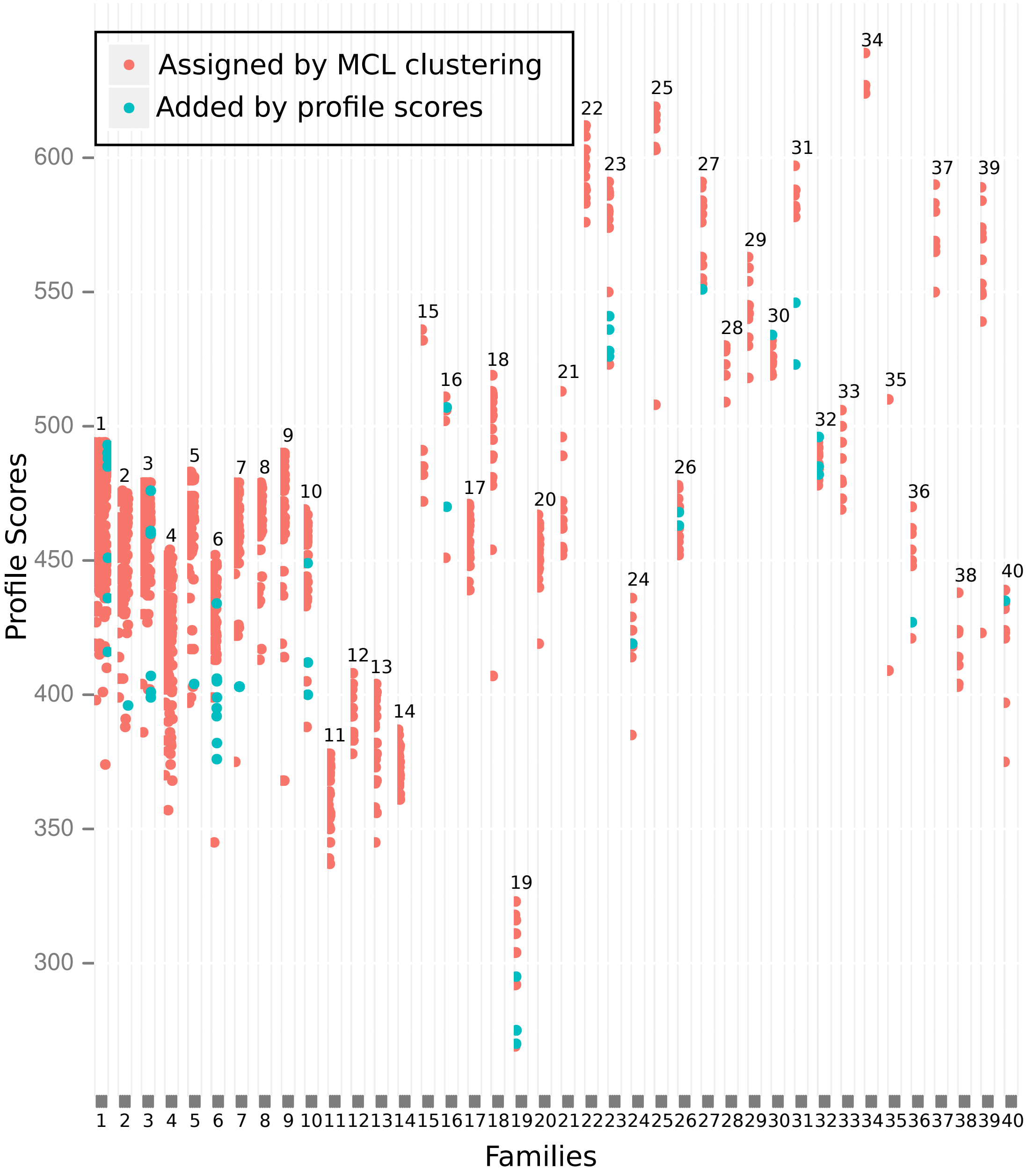


Figure D.3. Verifying repeat families with sequence profiles and re-assigning individual repeats. All repeats were clustered into families using Markov clustering [80, 320]. We verified these families using an independent method of sequence profiles, see Methods section "Clustering of repeat sequences into conserved sequence families". After the generation of one profile per family, we calculated the profile scores for each repeat in the `REPEATS` dataset. We plotted the profile scores (y-axis) for each repeat assigned to one of the families (x-axis) as red-coloured dots. Subsequently, we used this range of profile scores to re-assign repeats to one of the existing families as stated in the main text of the manuscript. Profile scores for re-assigned dots are in blue (73 repeats). These profile scores are also used to assign new input repeat sequences from the web server to one of our existing families. Figure taken from [P3].

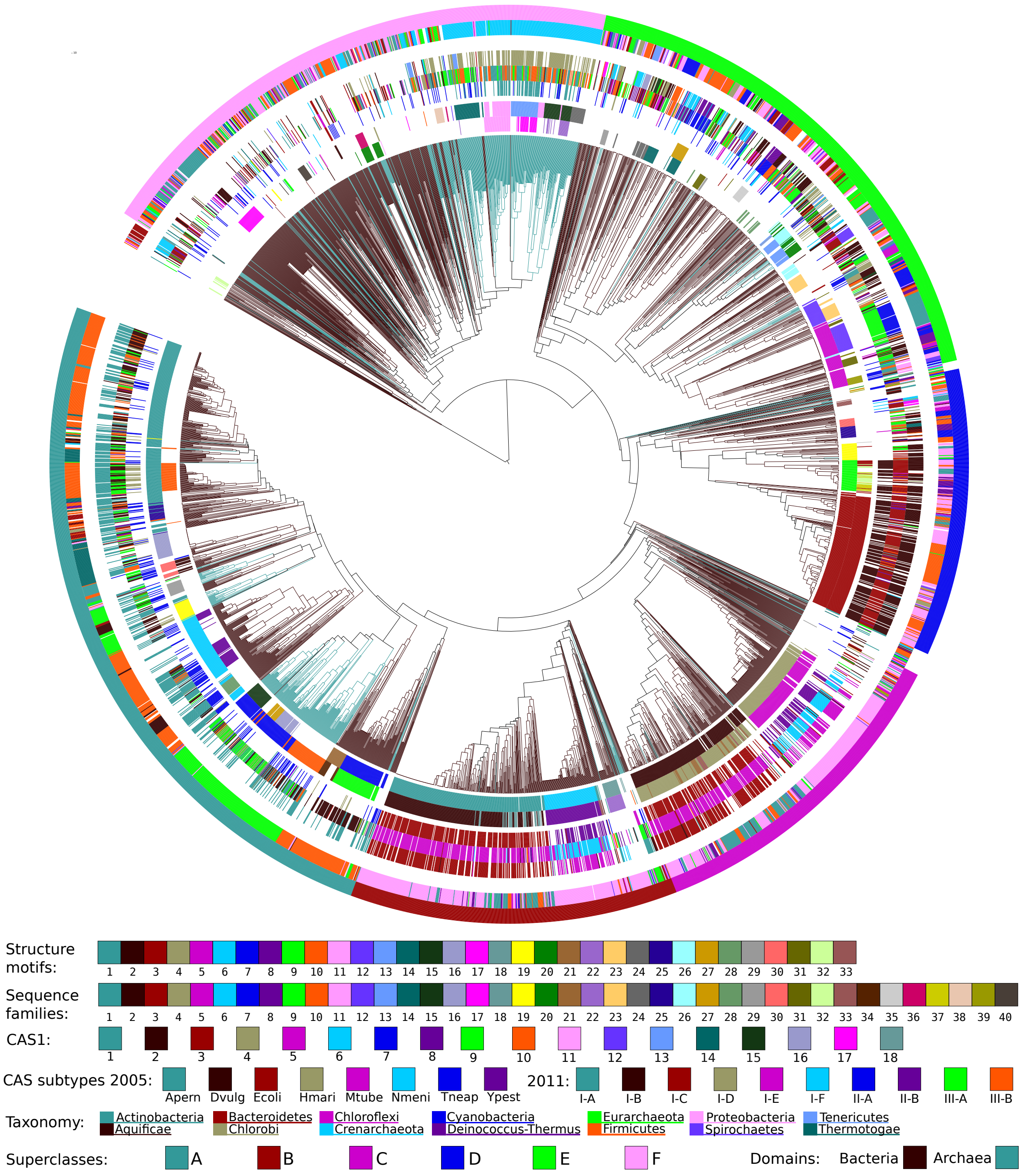


Figure D.4. CRISPR of repeat conservation including all annotations. CRISPR repeats cluster into 33 structure motifs and 40 sequence families. Here we show the cluster tree with all annotation rings—the “altogether option in the webserver—colour coding starts from inside to outside, see the legend. The branches of the tree are labelled according to the origin of the repeat: blue-green for archaea and dark brown for bacteria. **Ring 1** (inner-most) 33 structure motifs, **ring 2** 40 sequence families, **ring 3** Haft 2005 subtype annotation, **ring 4** Makarova 2011 subtype annotation, **ring 5** 18 cas1 clusters, **ring 6** taxonomic phyla annotation and **ring 7** (outer-most) the six superclasses for general orientation. Figure taken from [P3].

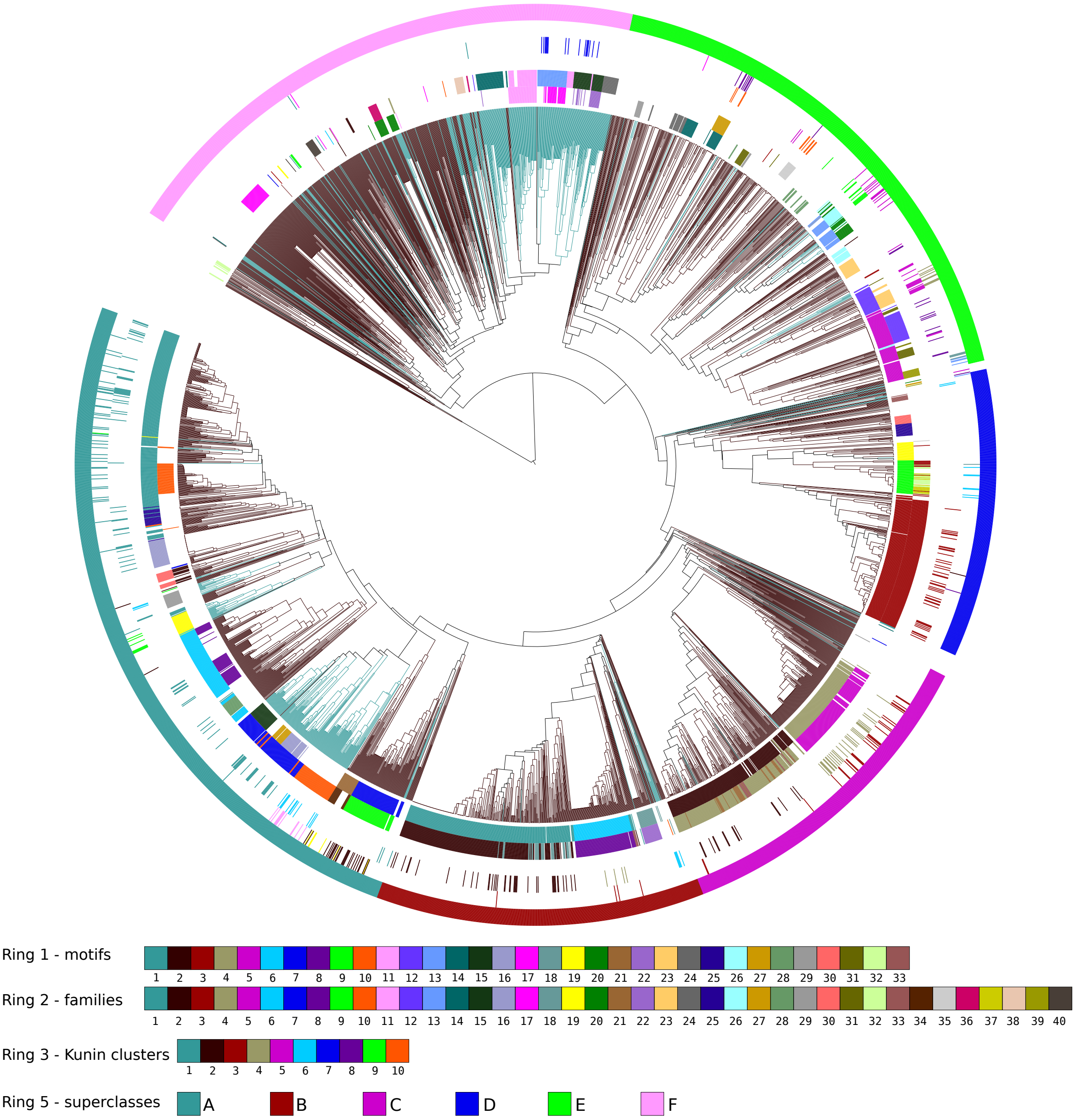


Figure D.5. Comparison of our clustering with previous domain-wide repeat clusters or families on our CRISPRmap tree. The branches of the tree are labelled according to the origin of the repeat: blue-green for archaea and dark brown for bacteria. **Ring 1** (inner-most) shows our structure motifs, **ring 2** shows our sequence families. After the white ring, we show ten of the twelve clusters from Kunin *et al.* [180, 287] in **Ring3**; clusters 11 and 12 contain fewer than ten repeats and to be consistent with our cluster minimum size, we have removed them here. **Ring 4** contains those sequences of the Rfam [95] database that are also contained in REPEATS (since we have all sequenced genomes to-date) and only families (16 out of 65) with at least ten sequences. We do not mark the family names here, but just want to show the relative locations of sequences in the CRISPRmap tree. **Ring 5** (outer-most) shows the six superclasses for general orientation. In summary, we clearly see that our data is significantly more comprehensive than previous work. Figure taken from [P3].

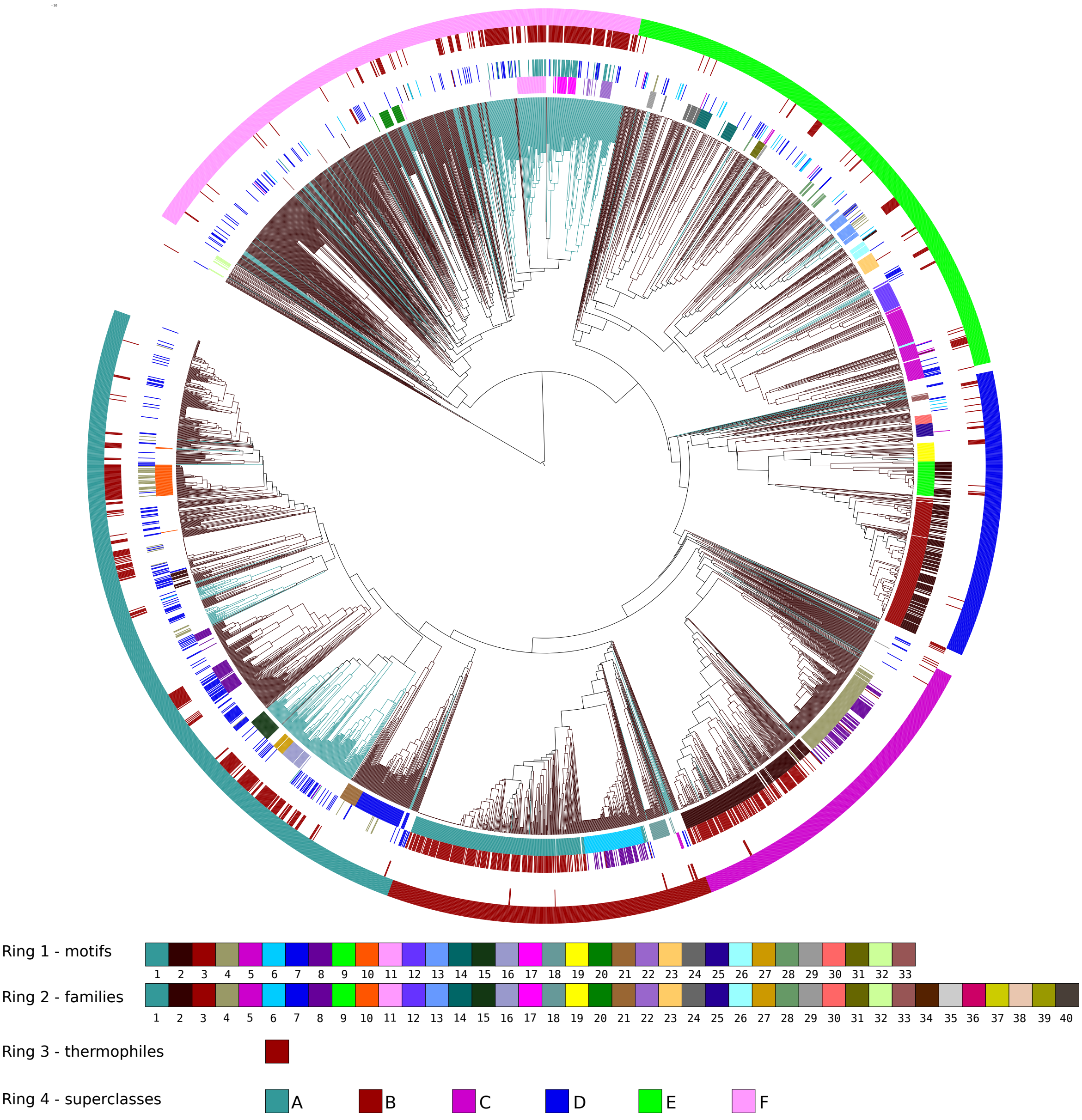


Figure D.6. CRISPRs found in thermophilic organisms. Ring 3 shows the number of CRISPRs that were found in thermophilic organisms (taken from ExtremeDB, `http://extrem.igib.res.in`, March 2013). At leat 17 % of our CRISPRs stem from thermophiles. Of these CRISPRs, 81 % are in superclasses A and F, which are associated with diverse types I-A, I-B, I-D, III-A and III-B. In contrast, only 7 % of the bacterial CRISPRs in superclasses B, C, and D—with strong Cas subtype associations—stem from thermophiles. The same is true for bacteria only: 60 % of the CRISPRs from bacterial thermophiles are in superclass A. Figure taken from [P3].

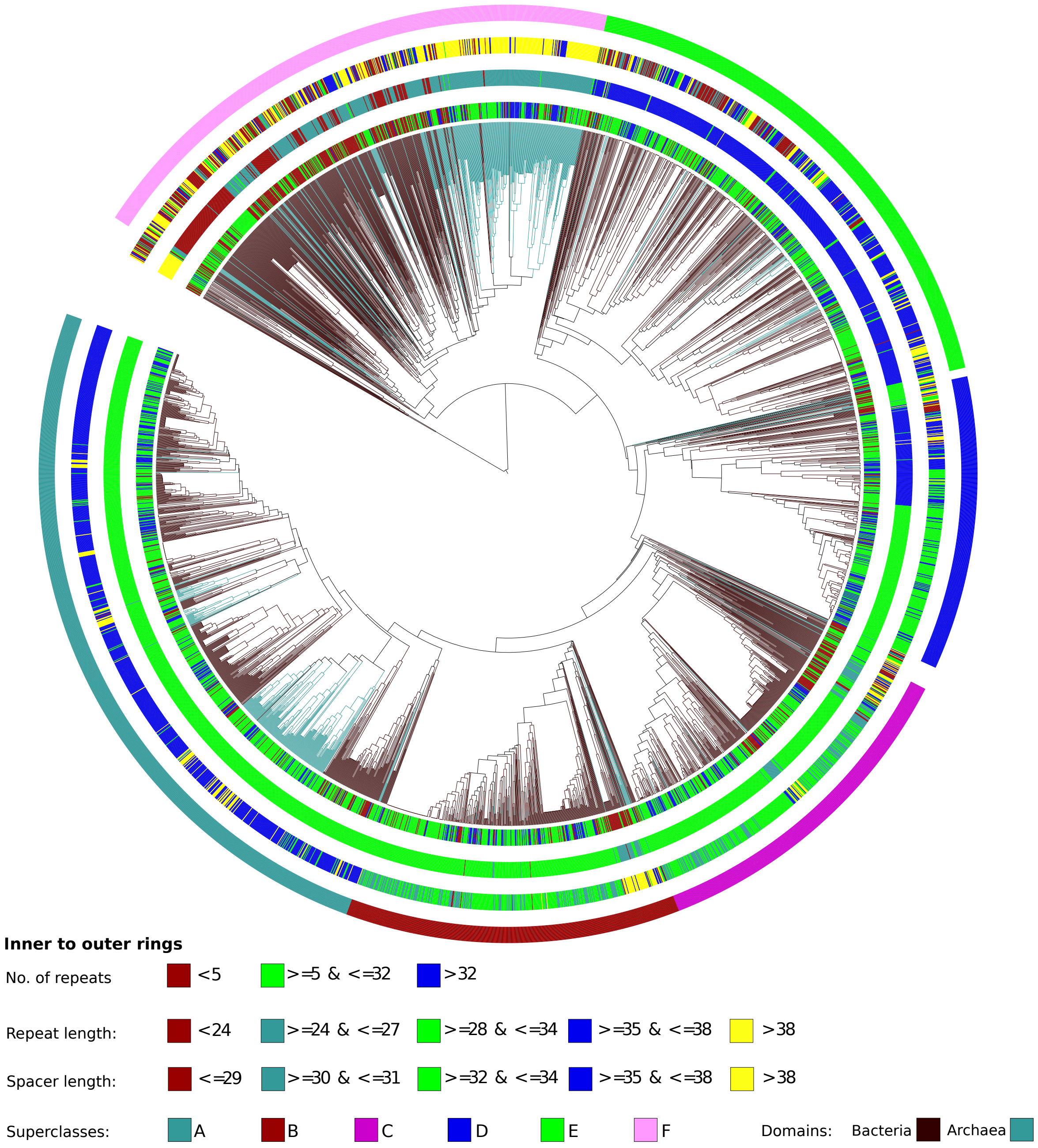


Figure D.7. Analysis of array, repeat and average spacer sizes. First, we see the very small arrays containing less than 5 repeat instances (red-brown) are mostly located in the more divergent parts of the `CRISPRmap` tree; most are within the bacterial part of superclass F. Many of these arrays may not be functional CRISPR-Cas systems, but other repetitive elements instead. Second, superclass F contains both some unusually short and unusually long repeats, which also may not represent functional CRISPRs. In addition repeats in superclass F and half of D are longer than those in superclasses A to the first half of D. Third, repeats in superclasses A and F are longer than ones in B-D; this means the Cas subtypes I-C, I-E, and I-F associate with shorter spacers than the others. Spacers in Crenarchaeota are unusually long with most longer than 38 nt. Interestingly, shorter repeats seem to pair with longer spacers. Cutoffs were chosen according to the distribution of each array characteristic. Figure taken from [P3].

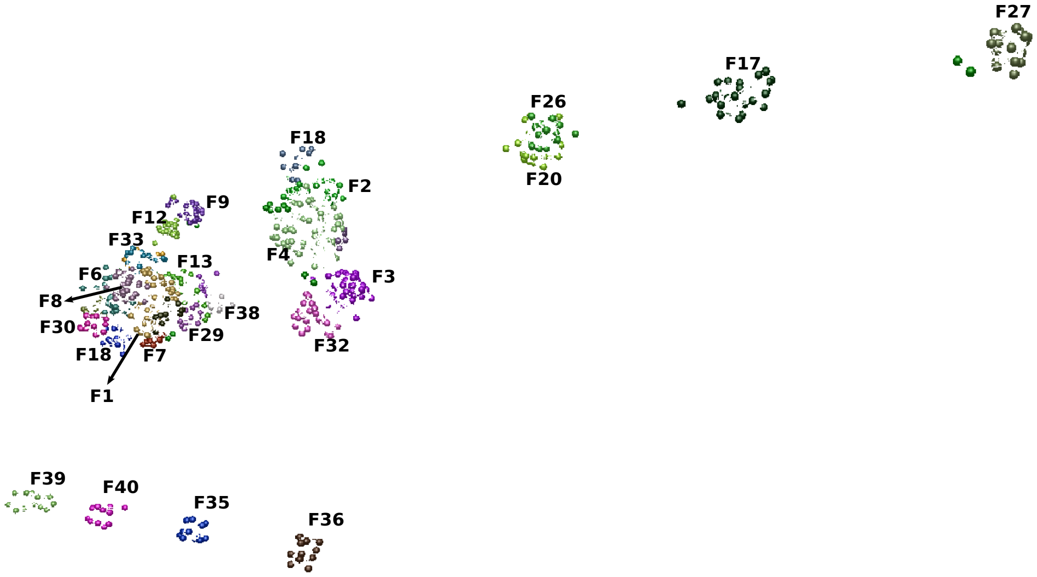


Figure D.8. Sequence families separated on a two-dimensional plane. The 40 sequence famlies are mapped onto a two-dimensional plane by BioLayout [309] according to their percent identity scores. We have marked only those families that are clearly visible. The families are divided into two main groups with some that are more separated from the rest. Figure taken from [P3].

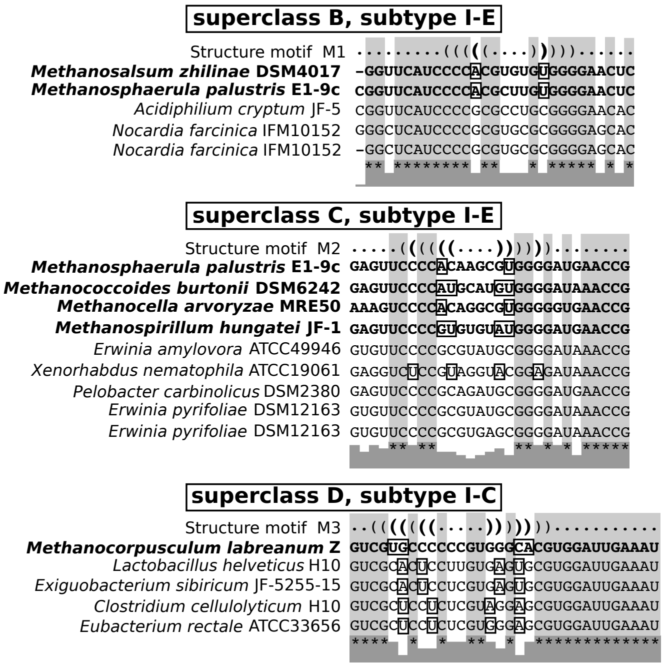

superclass B, subtype I-E
Structure motif M1 .........(((((....)))))......
Methanosalsum zhilinae DSM4017 –GGUUCAUCCCCACGUGUGUGGGGAACUC
Methanosphaerula palustris E1-9c CGGUUCAUCCCCACGCUUGUGGGGAACUC
Acidiphilium cryptum JF-5 CGGUUCAUCCCCGCGCCUGCGGGGAACAC
Nocardia farcinica IFM10152 GGGCUCAUCCCCGCGUGCGCGGGGAGCAC
Nocardia farcinica IFM10152 –GGCUCAUCCCCGCGUGCGCGGGGAGCAC
** ******** ** * ***** * *
superclass C, subtype I-E
Structure motif M2 .....((((((....))))))........
Methanosphaerula palustris E1-9c GAGUUCCCCACAAGCGUGGGGAUGAACCG
Methanococcoides burtonii DSM6242 GAGUUCCCCAUGCAUGUGGGGAUAAACCG
Methanocella arvoryzae MRE50 AAAGUCCCCACAGGCGUGGGGGUGAACCG
Methanospirillum hungatei JF-1 GAGUUCCCCGUGUGUAUGGGGAUGAACCG
Erwinia amylovora ATCC49946 GUGUUCCCCGCGUAUGCGGGGAUAAACCG
Xenorhabdus nematophila ATCC19061 GAGGUCUCCGUAGGUACGGAGAUAAACCG
Pelobacter carbinolicus DSM2380 GAGUUCCCCGCAGAUGCGGGGAUGAACCG
Erwinia pyrifoliae DSM12163 GUGUUCCCCGCGUAUGCGGGGAUAAACCG
Erwinia pyrifoliae DSM12163 GUGUUCCCCGCGUGAGCGGGGAUAAACCG
** ** ** * * *****
superclass D, subtype I-C
Structure motif M3 ..(((((((.....)))))))...........
Methanocorpusculum labreanum Z GUCGUGCCCCCCGUGGGCACGUGGAUUGAAAU
Lactobacillus helveticus H10 GUCGCACUCCUUGUGAGUGCGUGGAUUGAAAU
Exiguobacterium sibiricum JF-5255-15 GUCGCACUCCUCGUGAGUGCGUGGAUUGAAAU
Clostridium cellulolyticum H10 GUCGCUCCUCUCGUAGGAGCGUGGAUUGAAAU
Eubacterium rectale ATCC33656 GUCGCUCCUCUCGUGGGAGCGUGGAUUGAAAU
**** * * ** * *************

Figure D.9. Selected alignments showing evidence of horizontal transfer of structured CRISPRs from bacterial to archaeal genomes. Archaeal CRISPRs are indicated in **bold** typeface. The secondary structure from the respective motif is written above in dot-bracket format: brackets and dots corresponds to base pairs and unpaired nucleotides, respectively. The highlighted brackets and squares show that the secondary RNA structure has been conserved by compensatory base pair mutations. These compensatory base pair mutations give excellent evidence for the conservation and importance of the respective structure motifs. Figure taken from [P3].

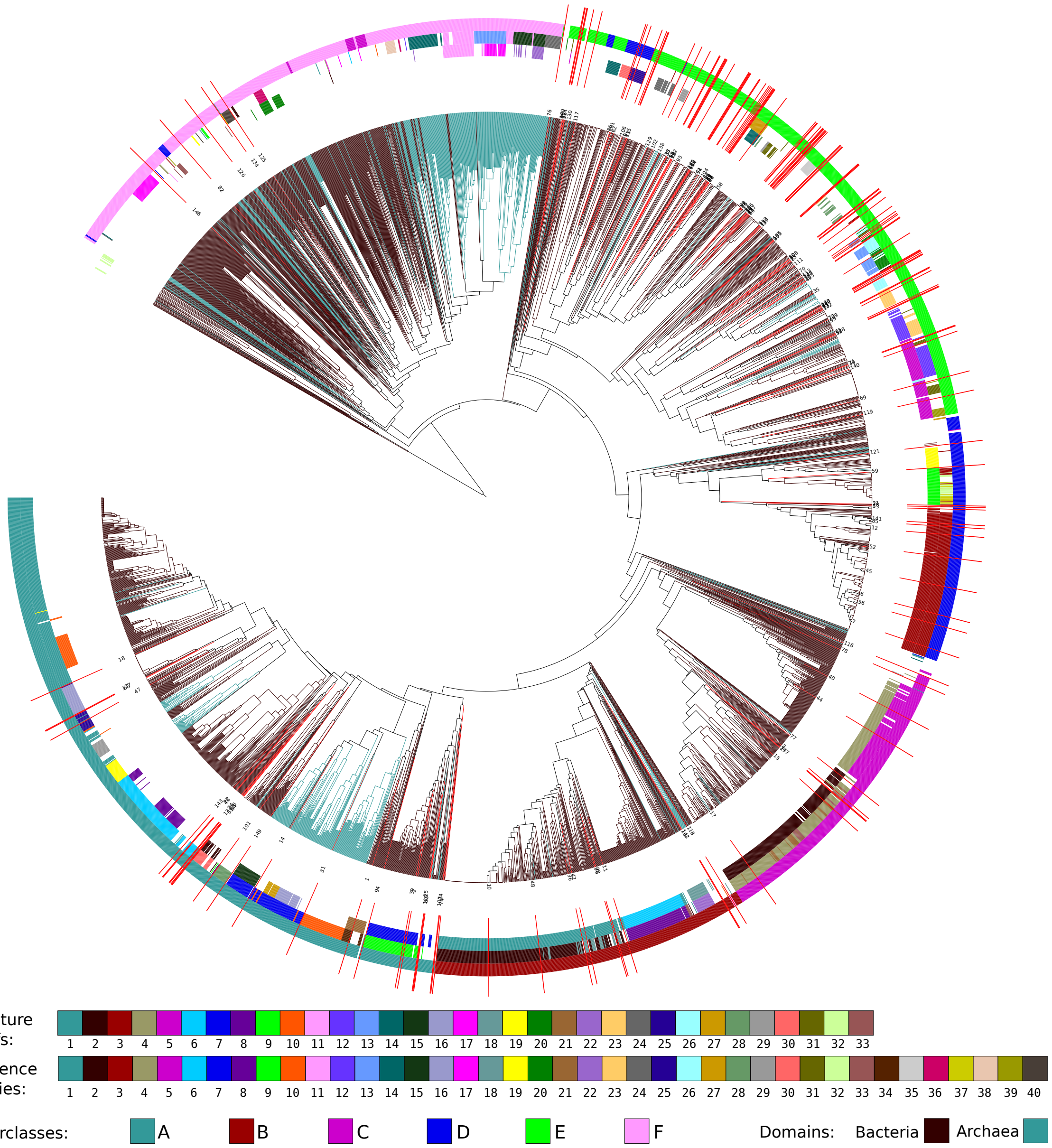


Figure D.10. `CRISPRmap` tree—a use-case study. This is the `CRISPRmap` cluster tree after re-clustering 150 repeats from a human metagenomic studies [256] together with our `REPEATS` data. The new 150 repeats are marked by red lines. Interestingly, many repeats have been assigned to superclass E and cluster together to potentially form new classes of motifs or families. Figure taken from [P3].

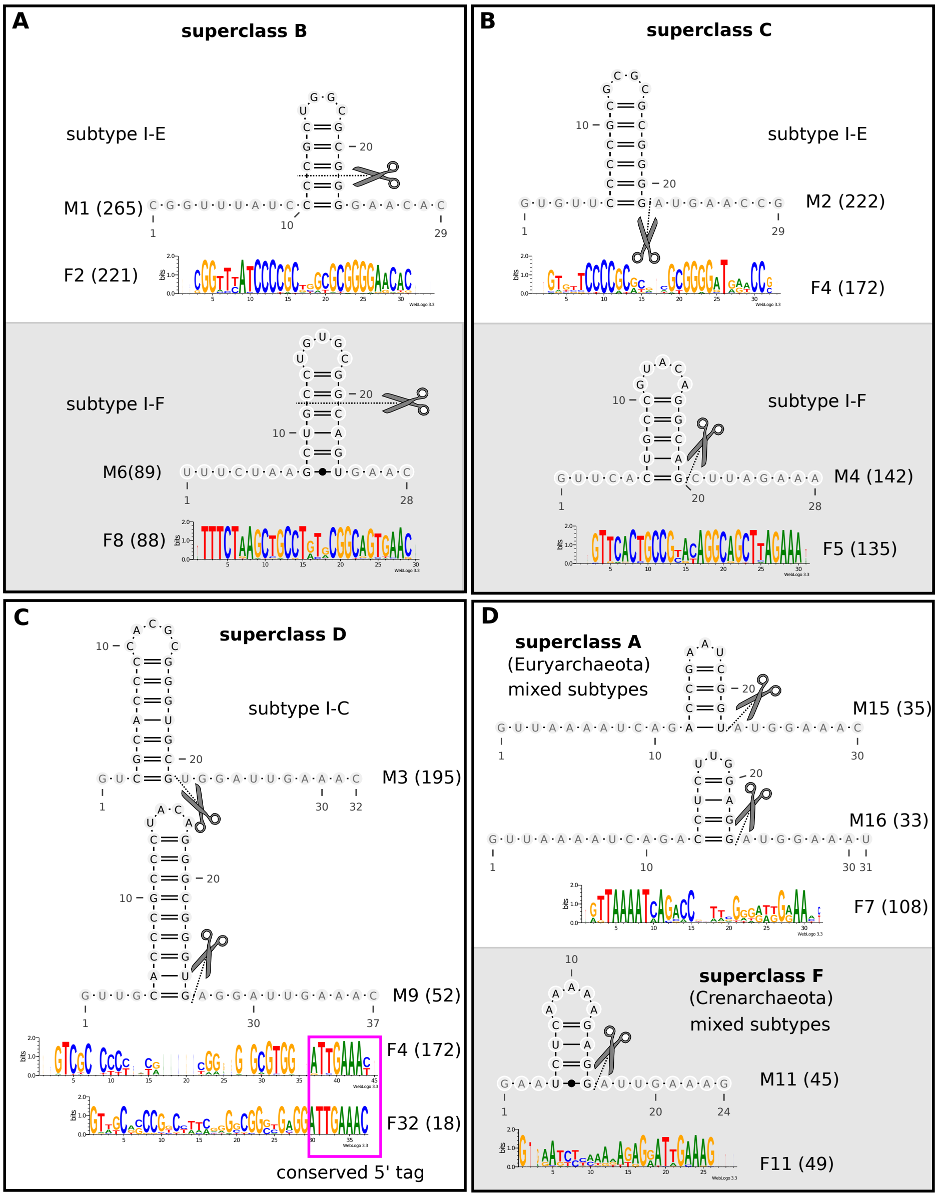


Figure D.11. Conserved structured CRISPRs fit well to published cleavage sites and display various patterns of sequence conservation. The sequence family logos correspond to the depicted structure motifs. Potential cleavage sites are indicated as observed in the literature [32, 99, 103, 129–131, 227, 248, 260, 274, 326]. (A)-(B) Superclasses B and C contain stable structure motifs of the subtypes I-E and I-F. The difference is that the structures in superclass B are closer to the 3' end of the repeat and that the potential cleavage site is in the double-stranded region of the stem instead of the 3' side of its base. (C) Superclass D contains members of the I-C subtype with relatively long hairpin motifs. Note that the potential cleavage site leads to an 11 nt instead of an 8 nt tag in the mature crRNA and we also see the well-conserved 3' end of the repeat (*ATTGAAAC*); this 3' sequence is found in many CRISPRs, also in archaea. (D) Examples of structure motifs found in archaeal repeats in superclasses A and F. These are smaller and less stable than the bacterial motifs. Figure taken from [P3].

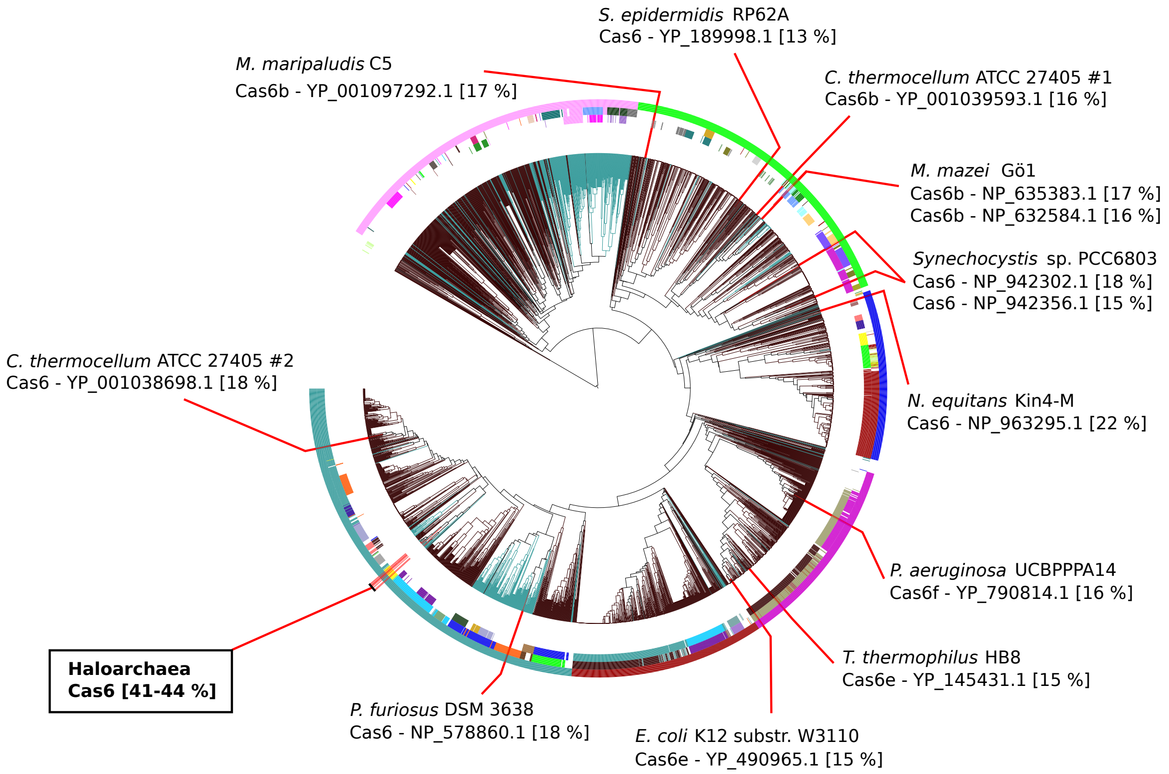

S. epidermidis RP62A
Cas6 - YP_189998.1 [13 %]
M. maripaludis C5
Cas6b - YP_001097292.1 [17 %]
C. thermocellum ATCC 27405 #1
Cas6b - YP_001039593.1 [16 %]
M. mazei Gö1
Cas6b - NP_635383.1 [17 %]
Cas6b - NP_632584.1 [16 %]
Synechocystis sp. PCC6803
Cas6 - NP_942302.1 [18 %]
Cas6 - NP_942356.1 [15 %]
C. thermocellum ATCC 27405 #2
Cas6 - YP_001038698.1 [18 %]
N. equitans Kin4-M
Cas6 - NP_963295.1 [22 %]
P. aeruginosa UCBPPPA14
Cas6f - YP_790814.1 [16 %]
Haloarchaea
Cas6 [41-44 %]
T. thermophilus HB8
Cas6e - YP_145431.1 [15 %]
P. furiosus DSM 3638
Cas6 - NP_578860.1 [18 %]
E. coli K12 substr. W3110
Cas6e - YP_490965.1 [15 %]

Figure D.12. The haloarchaeal CRISPR-Cas systems are different from other systems. The haloarchaeal CRISPR-Cas systems are distinct from published systems where the Cas6 protein has been functionally characterised. The circular hierarchical tree represents the sequence and structure similarity of repeats from all publicly available genomes, taken from the CRISPRmap web server [P3]. The locations of repeats associated with previously (partially) characterised Cas6 are highlighted with red lines: *Clostridium thermocellum* [260], *P. furiosus* [36,126,326], *E. coli* [32], *Thermus thermophilus* [103,155,274], *P. aeruginosa* [130], *Nanoarchaeum equitans* [248], *Synechocystis* [P8,P10], *Methanosarcina mazei* [P7], *Staphylococcus epidermidis* [129] and *Methanococcus maripaludis* [260]. The pairwise alignment percent identities in comparison with the Cas6 protein in H. volcanii are given in square brackets. For the CRISPRmap tree, brown branches represent CRISPRs from bacteria, the blue-green branches represent CRISPRs from archaea, the inner annotation circle represents different conserved structure motifs, the middle circle represents conserved sequence families and the outer circle represents the six superclasses. Figure and legend text taken from [P2].

D.2 Part III

Only some additional figures are provided for this part.

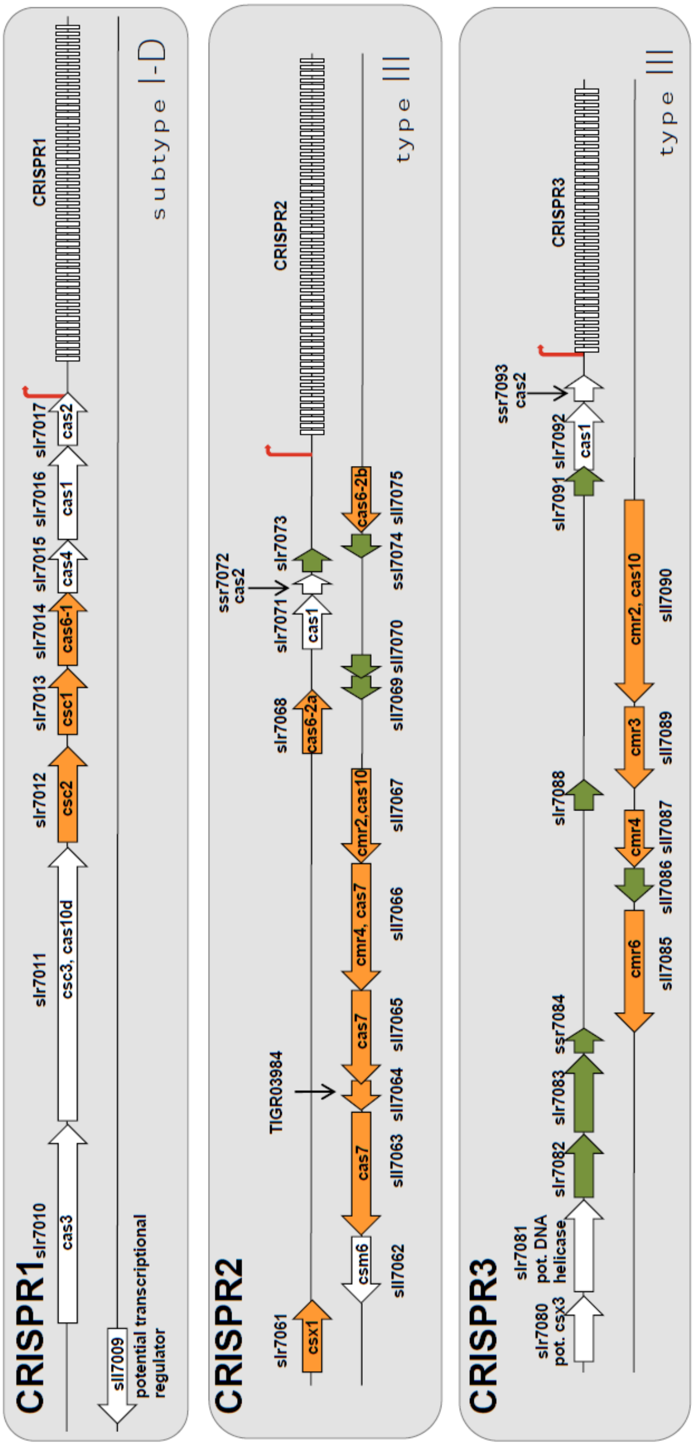


Figure D.13. Illustration of the CRISPR-*cas* loci CRISPR1–3. The pSYSA plasmid of *Synechocystis* sp. PCC6803 harbours three CRISPR-Cas systems, named CRISPR1–3. The CRISPR array and annotated *cas* genes associated with these arrays are depicted. Arrows in green represent genes coding for hypothetical proteins and arrows in orange illustrate cas-genes from the RAMP family. Experimentally mapped start sites of transcription (TSS) are marked by thin red arrows. Direct repeats are symbolised by narrow rectangles. For selected genes, we specify synonymous gene names. In general, however, we use the nomenclature introduced by Makarova and colleagues in [201, 202]. Figure taken from [P10].

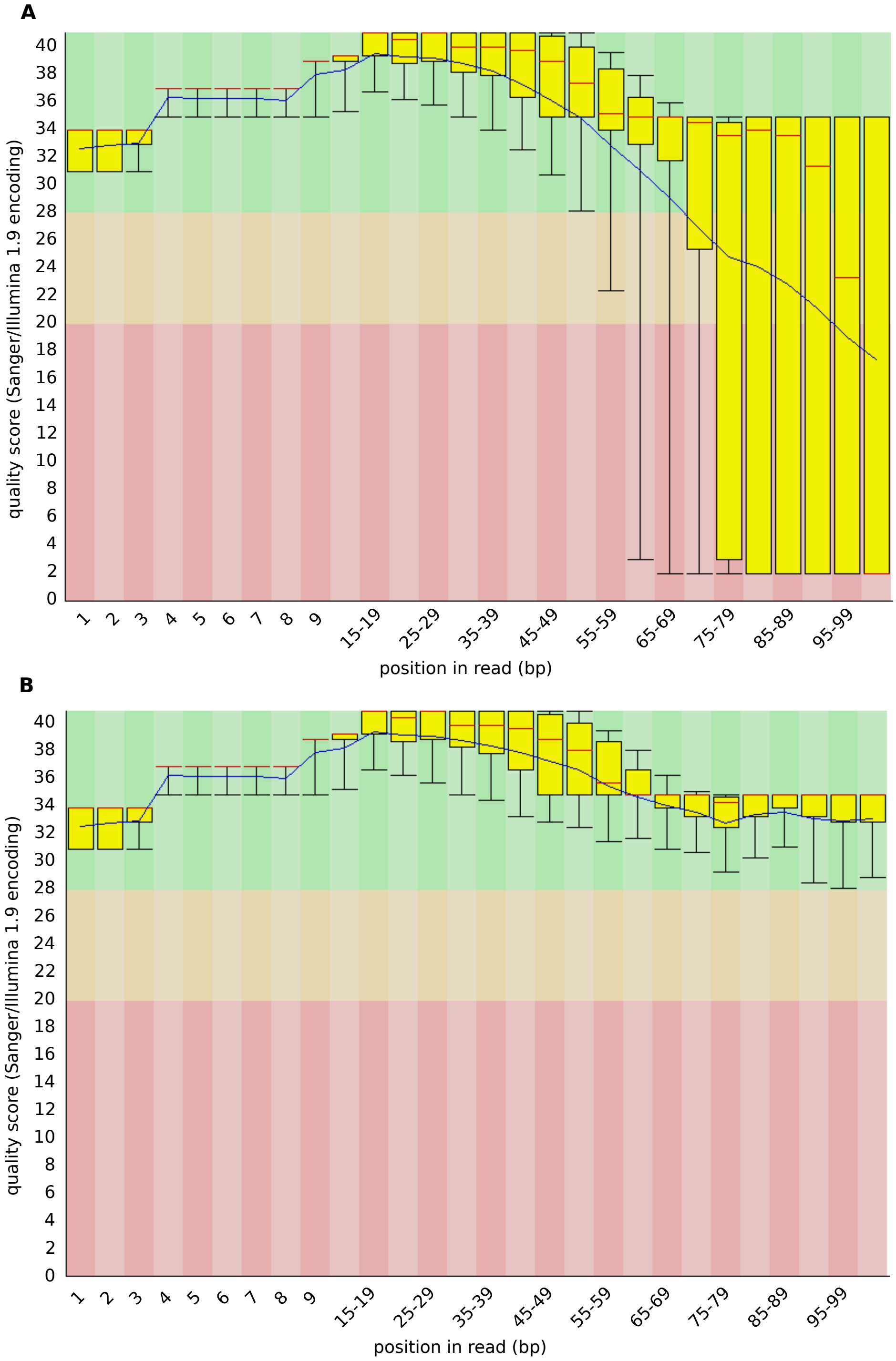


Figure D.14. Base-pair quality image from the `FASTQC` program for the `RNA-seq` dataset A. (A) We see an increasingly poor sequencing quality towards read ends for the original dataset, possibly due to the poly(A) tails and subsequent adapter sequences. (B) After quality trimming, we see that the read ends with a poor sequencing quality have been removed. Figure taken from [P10].

D.3 Part IV

The first half of this section is included in the supplement of [P4] and was included here for comprehensive purposes.

Algorithms for secondary structure prediction

We used `RNAfold` from the Vienna Package Version 1.8.4 as a representative of the global folding approach. The options used in this study are `RNAfold -d2 -p -noLP`. For folding under the constraint of the consensus structure, we used the additional option `-C`. `RNAfold` does not compute accessibilities, but position-wise accessibilities (as measured in the `YeastUnpaired` dataset and used in our evaluation) can be computed from the base-pair probabilities as defined in Equation 2.5 Section 2.5.6.

We use `Rfold` [170] in our analysis to represent this approach for base-pair probabilities and `Raccess` [171] for accessibilities. The commands for `Rfold` and `Raccess` were `run_rfold -max_pair_dist=L -print_prob=true` and `run_raccess -max_span=L -access_len=1`.

The execution call for `RNAplfold` is: `RNAplfold -noLP -W` W `-L` L `-u 1`.

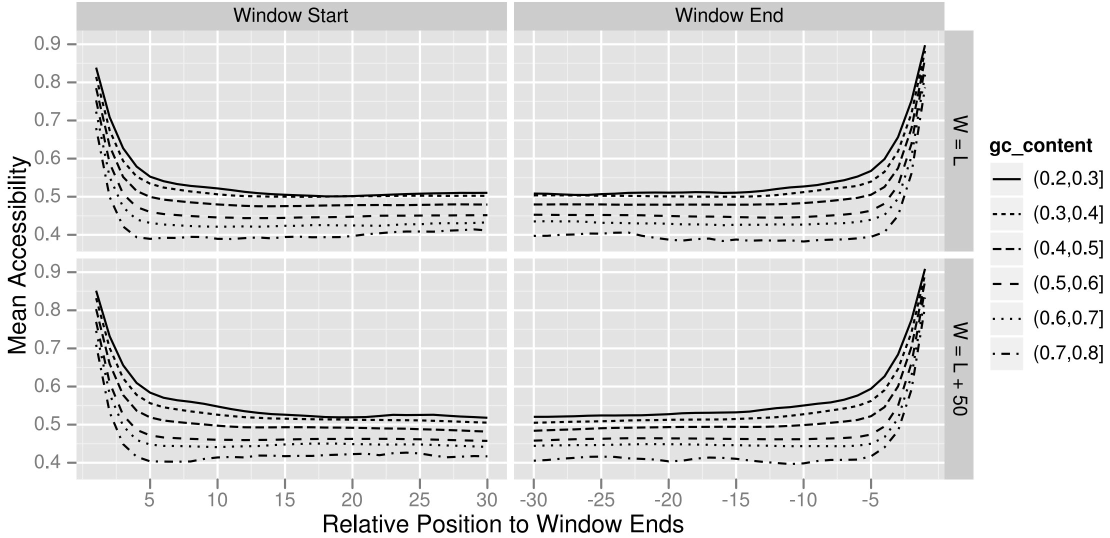


Figure D.15. Average accessibilities per window position for the 400 mRNAs used for Figure 6.3, split by *GC*-content of the windows. While average accessibilities decrease with increasing *GC*-content, border nucleotides are distinctly more accessible for all instances.

Curated benchmark set of *cis*-regulatory elements

From all 222 families labelled as "`Cis-reg`" in the `Rfam` database version 10.0 [95, 111], we have selected 98 with experimental evidence, which are likely to have well defined structures.

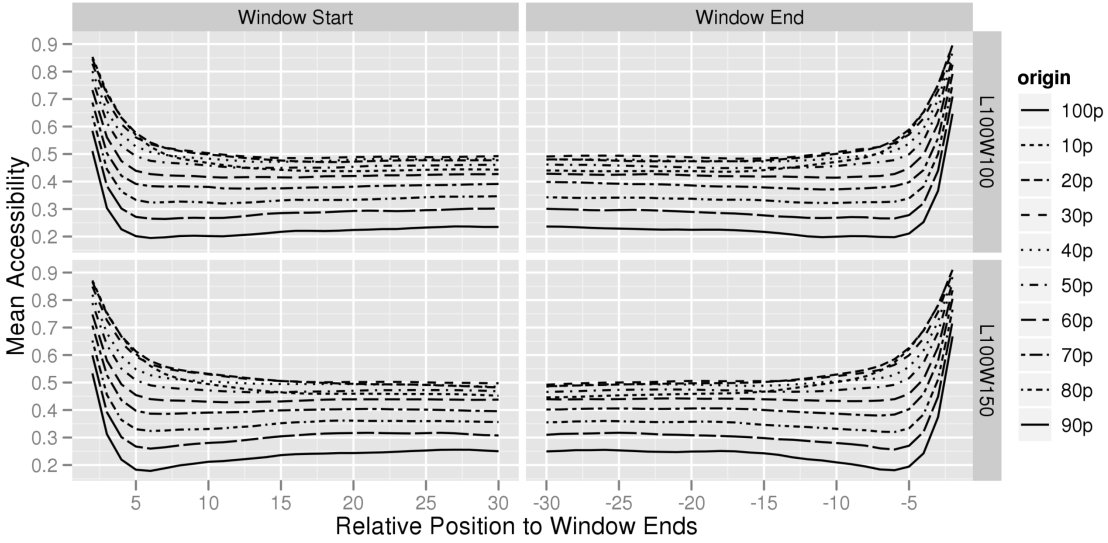


Figure D.16. Average accessibilities per window position for ten $15,000$ nt random sequences ranging in GC-content from 10–100 %. Sequences were folded with $L = 100$ and $W = 150$ (lower) and $W = 100$ (upper). Folding of each sequence resulted in $15,000 - W + 1$ independent folding windows.

These families comprise sequences from eukaryotic, bacterial and viral genomes with diverse *cis*-regulatory functions. More information about each one is available though the CisReg website, http://lancelot.otago.ac.nz/CisRegRNA/, with links to Rfam.

We extracted the seed alignments for each family from the database and used BLAST [5] to locate the positions of each element. Subsequently, we extracted the element sequences from the original sequences within contexts of 100, 200, 500 nt to either side of the functional element where possible. If there was not enough context, the sequences were extended to the beginning and/or end of the mRNA. We further extracted full-length mRNAs up to maximum context of 3,000 nt. Some of the original sequences are genomic from bacterial or viral genomes, so that possibly non-mRNA sequences are within the dataset. This fact, however, should not significantly influence the comparison of the prediction methods. We divided the dataset into originating from mRNA or genomic context to separately test the trends observed. To gain the exact base-pairs for each structural element in the family, we mapped the given consensus structure to the individual sequences. Any base-pair not consisting of GC, AU, or GU were omitted. Furthermore, any base-pairs that did not allow for at least three unpaired bases within a hairpin loop were also omitted.

The consensus structure only includes base-pairs that are common to all elements within a given family, although more base-pairs are likely to form in the individual element to improve its stability. To find the most stable structure, each element was folded with RNAfold using the consensus base-pairs as a constraint and the resulting minimum free energy structure was used. In the process of mapping the consensus structure to the single elements, non-conforming base-pairs were deleted. Therefore, we filtered out any elements that (1) did not retain at least 80 % of the original base-pairs in the consensus structure and (2) did not retain at least 80 % of the mapped base-pairs in the final constraint structure as folded by

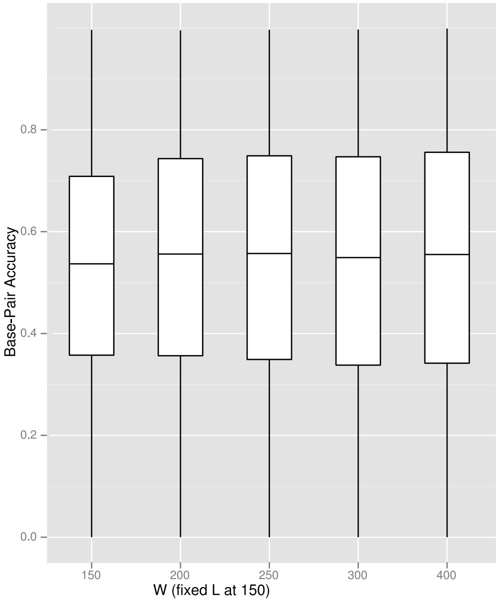


Figure D.17. Base-pair accuracy box-plots for the `CisReg` data for several window sizes W and span $L = 150$, using RNAplfold. For the comparison on the `CisReg` data shown in Figure 6.6, the span L was optimized using Rfold and thus independently from RNAplfold or LocalFold. Given this L, we selected the optimal W for RNAplfold. Results for $W = 200$ show the highest median accuracy and the smallest variance.

`RNAfold` (constraint folding with `RNAfold` sometimes results in a constraint base-pair being left unpaired – see `RNAfold` manual). This means, only sequences that were very similar to the consensus structure in `Rfam` were used.

The final dataset, referred to as `CisReg`, consists of 2500 individual structural elements from 95 *cis*-regulatory elements located on the mRNA across many different species. RF00632, RF00227, and RF00524 were not used because they did not pass the the filtering steps described above. Possible reasons for this are as follows: RF00632 (sxy 5' UTR element) includes only two sequences in the seed alignment and the 16 sequences in the full alignment from *H. influenzae* have 97 % identity, i.e., poor evidence for the consensus structure. RF00227 (ftz instability element 3') is mainly unstructured with a small nine base-pair stem in the centre. RF00524 (R2 RNA element) is a large computationally predicted structure that has a functional ribozyme within it. Subsequent updates to `CisReg` and `Rfam` entries should address these deficiencies. With this dataset, we evaluated over 85,000 base-pairs.

D.3.1 Dataset redundancy evaluation

During the manual curation the families were chosen to exclude similar families.

Similarity within families

To assess the sequence redundancy of our dataset, we took the sequences with 100 nt context to either side of the element. We selected this context, because the direct context is the most influential region for the structure prediction of the regulatory element. We subsequently

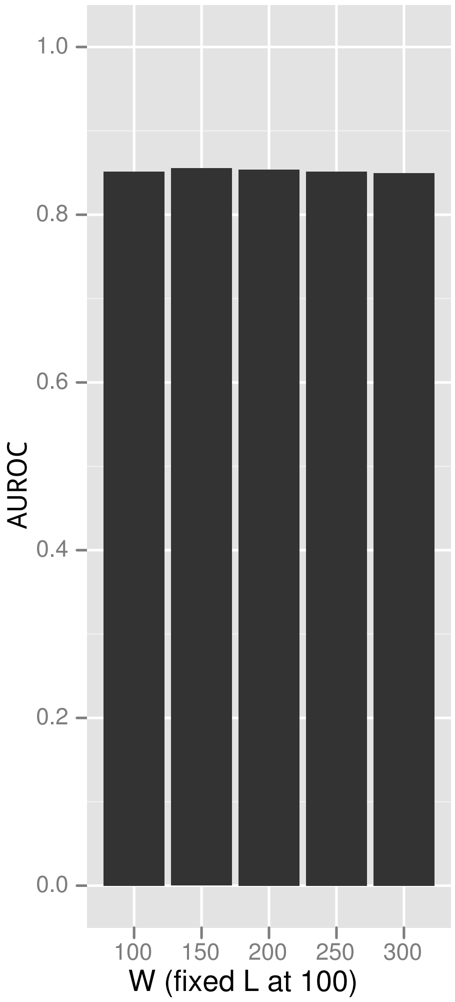


Figure D.18. Comparison of W values for a fixed L value on accessibility data AUROCs for separating high-scored and low-scored nucleotides from the `YeastUnpaired` data for several window sizes W and span $L = 100$, using RNAplfold. The comparison of `YeastUnpaired` in Figure 6.8 was done for several L (fixing W at $L + 50$). The best result for `RNAplfold` was reached using parameters $L = 100$ and $W = 150$. This is the optimal W for this span for `RNAplfold`.

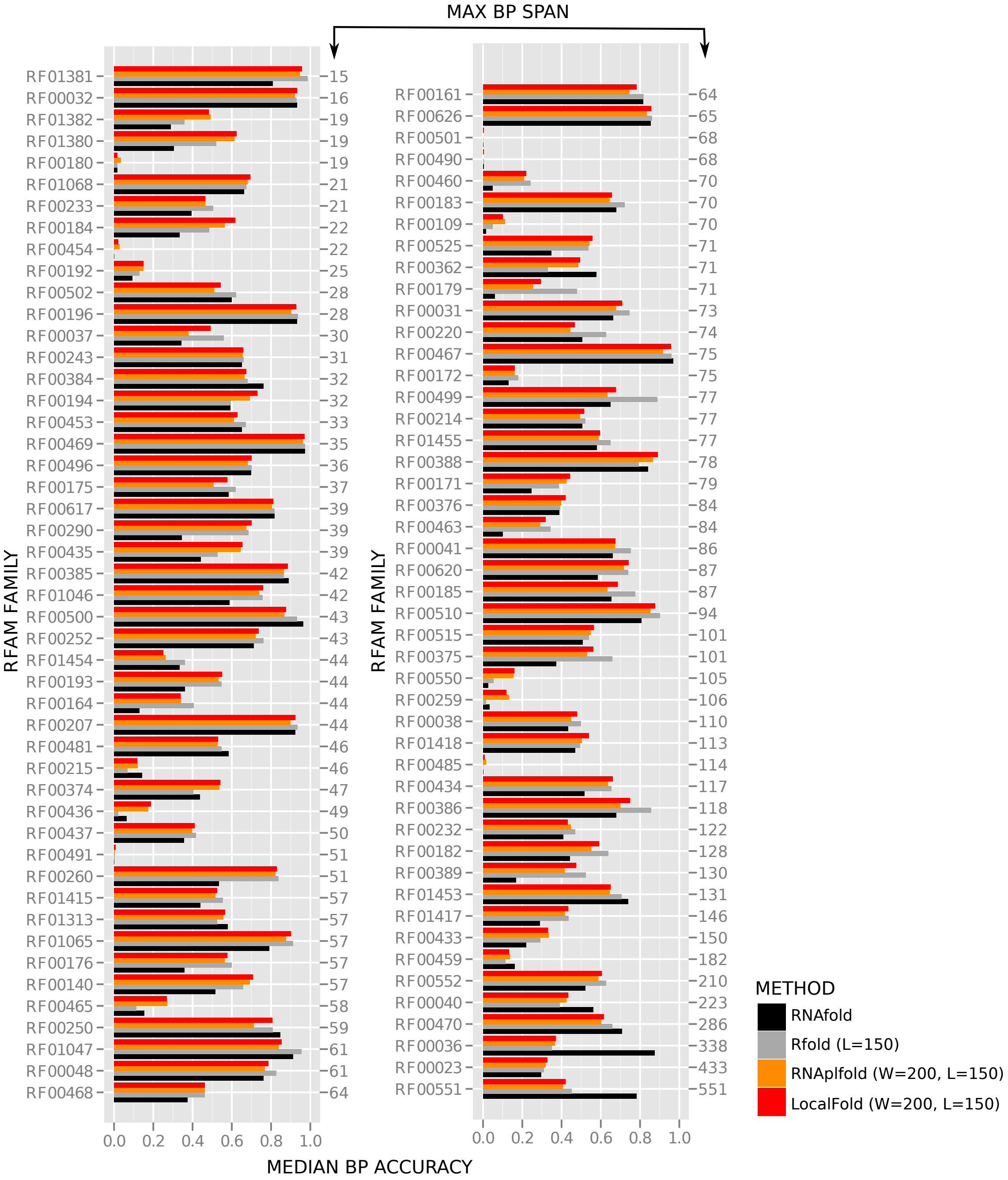


Figure D.19. Comparison of structure-prediction performances for individual `Rfam` families. The median *bp-accuracy* is shown separately for each of the 95 `Rfam` families within the `CisReg` data using sequence contexts of 500 nt. The families are sorted by the maximum base-pair span of their elements, ranging from 15 to 551. This information is more relevant than the actual element length, because this corresponds to the parameter L used. `RNAfold` only performs better than the other methods when the base-pair spans of the structure greatly exceeds the maximum base-pair span parameter $L = 150$. In general, we see similar trends across most families and no bias due to data redundancy is evident.

clustered these sequences using `BlastClust` [5]. This program groups sequences according to sequence similarity. To avoid the problem of overlapping sequences as described above, we set the coverage of both sequences to 100 %. To assess the amount of sequence similarity, we varied the percent identity of the pairwise alignments from 10 % to 100 % in steps of ten. We received the following number of clusters: 2,460 (100 % identity), 1,759 ($\geq$90 % identity), 1,671 ($\geq$10 to $\geq$80 % identity). Therefore, at least 1,671 sets are different problems with respect to structure prediction. Even a single mutation can alter the element structure at specific locations. Modest sequence differences (e.g. >20 %) usually result in different foldings and thus form different problem sets. As most of the redundancy is due to similar sequences within a family, we separated the analysis into families in Supplementary Figure D.19. Here we observe the same trends as we presented in the main paper for most of the families. In addition to the support of the `YeastUnpaired` results, this analysis exhibits the reproducibility of our results. The program call for `BlastClust` (pI = percent identity) was: `blastclust -i sequences_context100.fasta -o blastclust.out -p F -L 1 -b T -S pI`. We have also reported the overall pairwise similarities of the seed alignments in the online database.

Similarity between families

The clustering analysis on primary sequence indicate that there are distinct sets, however to access the redundancy in folds we have used `cmcompare` [142] to do pairwise comparisons of each of the covariation models to one another and to the whole `Rfam` 10 set of models. For cmcompare scores of over 20 are considered 'worthy of note' and 7.4 % of the entire `Rfam` database had such scores. However, no pairs within the `CisReg` set used here had scores over 20. Although there were some notable matches to other `Rfam` families not in the set analysed here. We also compared the primary sequence of the first member of each family to all the sequences in the other families using blastn. Only 9 of 98 had matches with $E < 1.0$ in the other sequences. These were all short regions of identity < 13 bases long, too short to contribute substantial common secondary structures.

D.3.2 Sequences used to evaulate cleavage events of single repeat instances in CRISPR2

The nine *in-vitro* experiments to prove or disprove cleavage at single repeat instances within varying sequence contexts were performed using the following sequences. Structure prediction to calculate structure accuracies was also performed on the full sequences separately. The descriptive identifier is built up as follows: experiment number, specific spacer and repeats and the length of the entire fragment, separated by underscores. If a spacer or repeat is not in full length, then the number of nucleotides is given after the respective part. All sequences begin with `GG` and are separated in the table by repeat instances, i.e., the rows alternate between spacer and repeat sequence parts.

Fragment	Sequence
I	`GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCCCGUGGUGGGAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUAA`
II	`GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCCCGUGGUGGGAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUUAGGCA`
III	`GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCCCGUGGUGGGAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUUAGGCAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGAGCGG`
IV	`GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCCCGUGGUGGGAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUUAGGCAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGAGCGGGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCGUUAUCCGGCAAAGAAACCACACUACUAAGCUCGACAA`
V	`GGUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUUAGGCAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGAGCGGGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCGUUAUCCGGCAAAGAAACCACACUACUAAGCUCGACAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGGGCCGGGCGCGAGUUGUCCUCCUGUCCGAGGCCCCAC`
VI	`GGCCCCGCCCCCGUGGUGGGAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUUAGGCAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGAGCGGGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCGUUAUCCGGCAAAGAAACCACACUACUAAGCUCGACAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGGGCCGGGCGCGAGUUGUCCUCCUGUCCGAG`

VII	GGCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUUAGGCAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGAGCGGGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCGUUAUCCGGCAAAGAAACCACACUACUAAGCUCGACAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGGGCCGGGCGCGAGUUGUCCUCCUGUCCGAGGCCCCAC
VIII	GGUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGAGCGGGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACCGUUAUCCGGCAAAGAAACCACACUACUAAGCUCGACAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGGGCCGGGCGCGAGUUGUCCUCCUGUCCGAGGCCCCAC
IX	GGCGUUAUCCGGCAAAGAAACCACACUACUAAGCUCGACAAGUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAACUGGGCCGGGCGCGAGUUGUCCUCCUGUCCGAGGCCCCAC

D.3.3 Dotplots showing the differences in base-pair probabilities between cleaved and uncleaved repeat loci

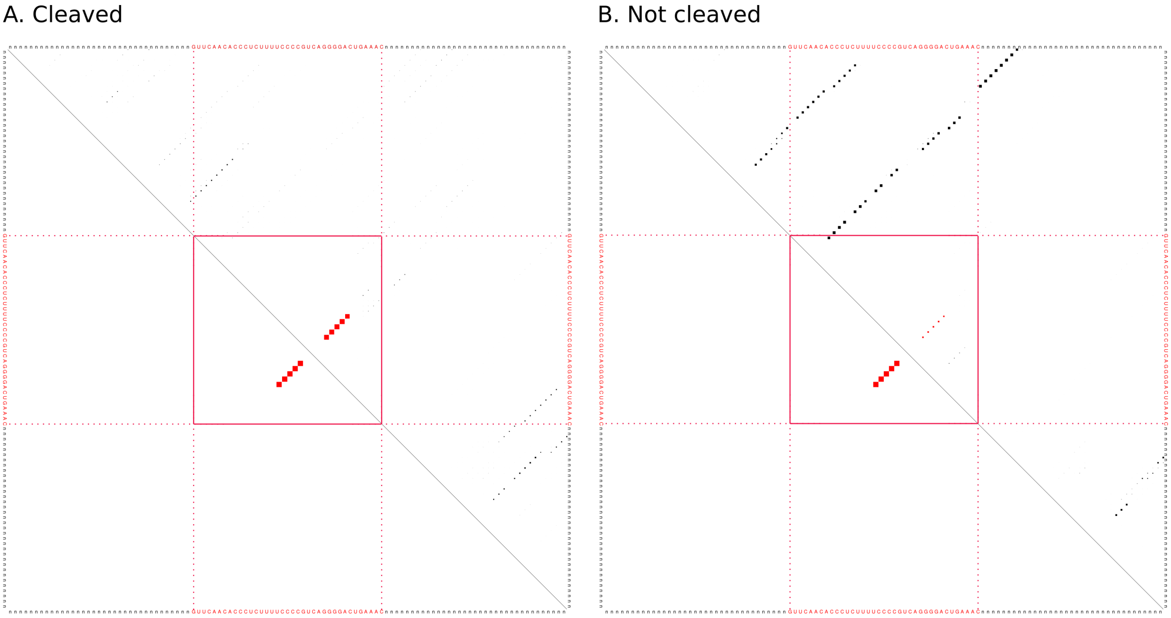


Figure D.20. Dotplots showing the differences in base-pair probabilities between cleaved (A) and uncleaved (B) repeat loci. Each dotplot depicts the average base-pair probability for each repeat instance from Section 7.2 in the upper right triangle and the structure motif (base-pairs highlighed as red dots) for CRISPR2 in *Synechocistis* sp. PCC6803 in the lower left triangle. Base-pairs between two nucleotides in the repeat are within the red box; base-pairs between a nucleotide in the repeat and one in the spacer are between the two dotted lines; and base-pairs between to nucleotides in the spacers are in the outer boxes. The spacer sequences are variable, depending on the repeat locus and we calculated average base-pair probabilities for the mode spacer length.

D.4 Part V

D.4.1 Additional data and results relating to the significance of accessibility around RRE sites in the RNAi pathway

We gathered five datasets containing RNAi interaction sites for siRNA and miRNA from three different organisms, which are phylogenetically very distant: humans, firefly and *Arabidopsis thaliana*.

Interaction data for endogenous miRNA

The endogenous miRNA data is divided into functional and non-functional interactions.

- `01-AtmiR:` This dataset was previously described in Section 8.1. The description is extended here by the GEO accession numbers from which the data was derived in Table D.21. It is a high-fidelity dataset of 110 functional and 114 non-functional miRNA-MRE interactions in *A. thaliana*.
- `02-Human:` This dataset consists of 67 functional MRE sites in 36 mRNAs of *Homo sapiens* (human) taken from `miRecords` (extraced from http://mirecords.biolead.org/download.php, release 5 May 2010). Entries in miRecord were only taken if mutation experiments were performed and the target site could be located in the given mRNA accession number [346]. Since no non-functional data was available here, a corresponding set of random MRE sites was generated as follows: for each functional MRE, a second, non-overlapping region was extracted from the same target mRNA of the same length. Although by chance one could hit an "unknown" but functional MRE, in general the random sites can be assumed to be non functional. The final dataset contains 67 functional and 67 non-functional MREs.

Interaction data for artificial siRNA

Artificial siRNAs are designed to knock down the expression of target genes by binding to their corresponding mRNAs, analagously to the artificial miRNAs described in Chapter 10. For the following datasets of binding sites of artificial siRNAs, repression efficiency scores are available. Thus, we do not have a partitioning into functional and non-functional sites, but continuous values of repression efficiency instead. All three datasets were extracted from [303]. First, the knock-down efficiency of each siRNA on target mRNA was measured by the average mRNA repression. Let X be the set of measurements for all siRNA-target pairs. Then, all $x \in X$ were subsequently normalised to values between $[0, 1]$ with a linear interpolation:

$$f(x) = \frac{x - min(X)}{max(X) - min(X)},$$

where $max(X)$ and $min(X)$ are the highest and lowest repression measurements for all observations in the set X, respectively.

- 03-Tafer02: consists of artificial siRNAs that target arbitrary regions of the coding sequences of human mRNAs MAP2K1, GAPDH, PPIB, and LMNA. It comprises 294 interactions in total.
- 04-Tafer03: consists of an independent set of artificial siRNAs that also target human mRNAs, in this case Cyclophilin, ALPPL2 and DBI. This set comprises 270 interaction sites.
- 05-Firefly: contains measurements of the repression efficiency of 89 artificial siRNA interactions, targeting only the luciferase mRNA in the firefly (*Photinus pyralis*).

Table D.21. GEO accession numbers for expression data of *A. thaliana genes* from ASRP [120]

Small RNA 454 sequencing	
col-0	GSM154336
dcl1-7	GSM154361
Gene expression micorarrays	
col-0	GSM47011, GSM47012, GSM47013, GSM47020, GSM47021, GSM47022, GSM47049, GSM47050, GSM47051
dcl1-7	GSM47023, GSM47024, GSM47025, GSM47026, GSM47027

Accessibilities in the vicinity of RNAi-regulatory recognition sites

Using a sliding-window approach as introduced in Section 8.2, we performed similar experiments with all datasets described above and plotted them in Figure D.21. For the sets of artificial siRNAs, we calculated Spearman's ranked correlations with repression efficiency scores, instead of performing two-sample tests to determine the differences of distributions between functional and non-functional interactions. In all datasets, functional sites are always close to regions of higher accessibility, usually downstream of the binding sites. Although the organisms and the exact location of these accessible regions differ, it is remarkable that they all show such patterns. In particular, it is of special interest that the regulatory recognition elements (RREs) of both miRNAs and siRNAs are not accessible, despite the common belief that these should be more accessible [213, 215, 305, 310, 342].

D.4.2 Deriving high-fidelity miRNA-MRE interaction data from CLIP-seq experiments in humans

Scanning for seed interactions in target mRNAs

Since calculating hybrid interactions between miRNA and MRE is computationally expensive, an efficient seed scanner was devised to locate potential sites of interactions. These sites were subsequently used as anchors to predict the extended hybrid pattern.

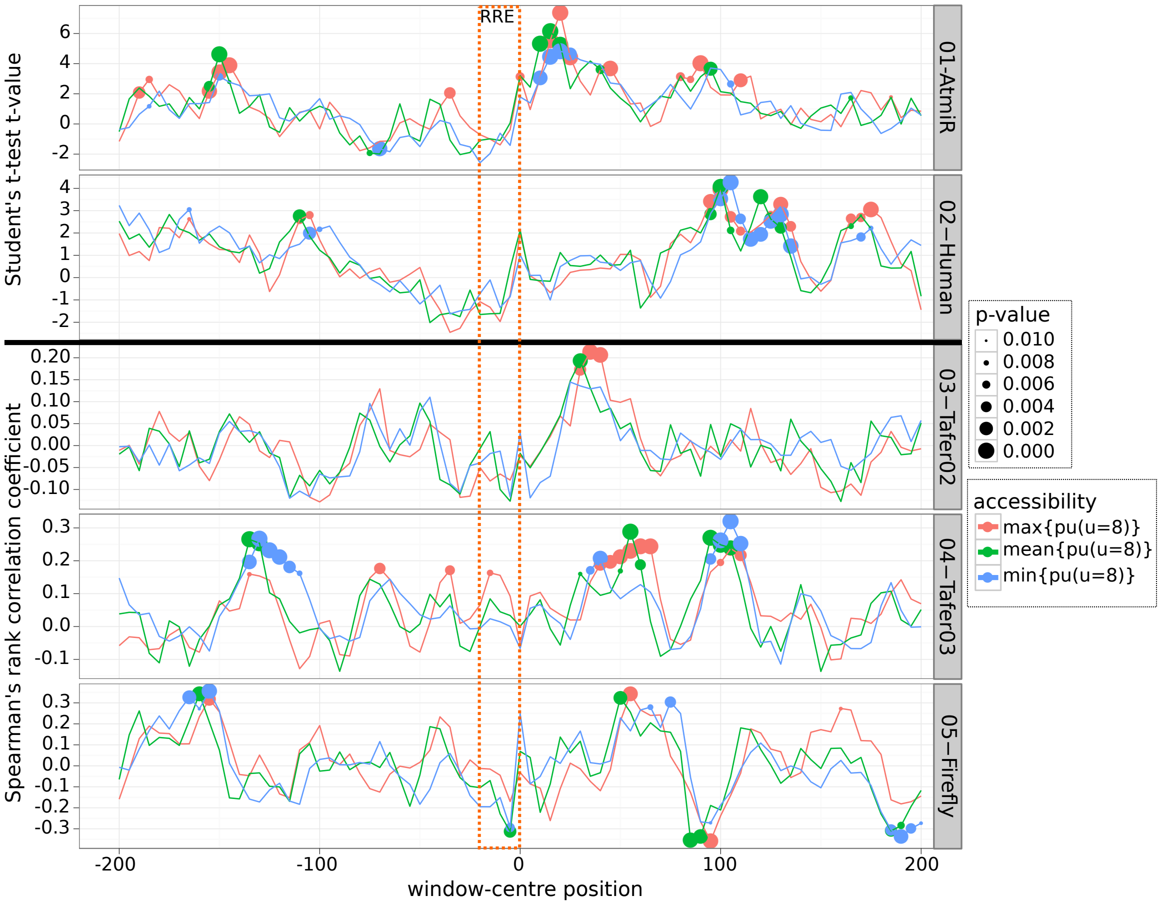


Figure D.21. Accessibility is significantly higher generally in regions downstream of target sites in the RNAi pathway. The same analysis performed in Section 8.2 was repeated for five datasets of miRNA and siRNA binding in *A. thaliana* (`01-AtmiR`), humans (`02-Human, 03-Tafer02, 04-Tafer03`) and in fireflies (`05-Firefly`); see further descriptions above of the data. For each independent dataset, the centre of 20-nt windows is indicated on the y-axis and either the t-value for Student's two-sample t-test for binary data or the Spearman's ranked correlation for continuous affinity measurements are given on the x-axis. The differently sized dots represent the p-value. Instead of the Student's t-test p-value, we used an independent test for calculating p-values for binary data—the Wilcoxon Rank Sum test.

The canonical seed definition is that the target MRE is complementary to positions 2–7 of the miRNA, excluding *GU* base pairs. However, in additional to the canonical seed, many interactions with non-canonical seeds have shown a regulatory effect in the literature [45, 65, 121, 196, 216, 306, 306]. Non-canonical seed interactions within positions 2–7 of the miRNA include *GU* base pairs [65, 187, 196, 216, 306], bulges in the miRNA [121, 216], bulges in the mRNA [45, 306] and further mismatches between miRNA and MRE [196, 306]. In addition, in a computational study on the stability of miRNA-MRE hybrids bound by Argonaute proteins, it was concluded that multiple *GU* pase pairs and bulges with single nucleotides at several positions in both sequences do not affect the overall stability of the interaction [345]. We have used the supplied evidence in the literature to define an extended seed interaction as follows:

1. Contains at least six base pairs between positions 1–8 of the miRNA.
2. Contains an arbitrary number of *GU* base pairs.
3. Contains a single-nucleotide bulge between positions 2–8 of the miRNA in the target MRE sequence.
4. Contains a single-nucleotide bulge between positions 3–8 of the miRNA in the miRNA sequence.
5. Contains at most one mismatch (internal loop with single unpaired nucleotides) within the seed region of 1–8 of the miRNA in both sequences.

The last three properties are mutually exclusive so that only one of the conditions (3)–(5) may apply. For each of the mature miRNA sequences being analysed, the seed sequences is extracted and complementary seed interaction sites are scanned in target RNA sequences. Seed scanning was performed using regular expressions. The previous seed definition is subsequently divided into the following seed types and scanning proceeds hierarchically, starting from the most extensive and ending in the most loose seed type:

1. **8-mer seeds:** contain base pairing between all *eight* miRNA positions 1–8 and the MRE, including *GU* base pairs.
2. **7-mer seeds:** contain *seven* consecutive base pairs between miRNA and MRE within positions 1–8 of the miRNA, including *GU* base pairs.
3. **6-mer seeds:** contain *six* consecutive base pairs between miRNA and MRE within positions 1–8 of the miRNA, including *GU* base pairs.
4. **Loose non-canonical seeds:** contain seed interactions that remain according the previous definition and at least six, non-consecutive base pairs between positions 1–8 of the miRNA.

Positions of seed interactions in target mRNA sequences are stored in BED format[1]. Applying `intersectBed` from the software package `BEDTools`[2], the intersection of the resulting seed types was used to filter overlapping seed interactions; the strongest seed type was retained.

Curation of functional miRNA-MRE interactions

Functional miRNA-MRE interactions were generated in a two-step process from `CLIP-seq` experiments where cross-linking with Argonaute proteins was considered evidence of miRNA binding: (1) locating and extracting the RREs of the Argonaute proteins with sufficient sequence context; and (2) the calculation of probable base-pairing patterns with expressed miRNA sequences.

AGO-RRE sites were derived derived from a `CLIP-seq` experiment performed in human embryonic kidney (HEK293) cells using the `PAR-CLIP` protocol cross-linking Argonaute proteins 1 to 4 (AGO1–4) to bound RNA [122]. The supplementary material of the publication provides 17,319 RREs bound by AGO1–4—according to the experiment. The `PAR-CLIP` protocol ends with an `RNA-seq` experiment using the RNA bound to the selected protein. A consequence of the `PAR-CLIP` cross-linking procedure is that the subsequent copy DNA carries T to C mutations at the cross-linked sites. Hence, a region was considered to be an RRE when a cluster of at least 5 overlapping reads, mapped to the human genome assembly hg18, contained a minimum of 20 % T to C mutations. An RRE sequence of 41 nt was extracted, centered on the position with the most frequent T to C mutations. Since we required the context sequence of the RREs for subsequent encoding of miRNA-MRE interactions, we had to determine the exact location of the RREs within their native mRNAs. First, the set of all mRNA transcripts from the NCBI RefSeq database were downloaded from the UCSC genome browser[3] in April 2013. This set contained 34,038 mRNAs from the more recent human genome assembly hg19. RRE sequences taken from [122] were aligned with a locally set-up `BLAST` [5] database of the downloaded mRNA transcripts using the nucleotide–nucleotide `BLAST` tool, `blastn`, version 2.2.25+; the E-value threshold was set to 0.0001, the DUST filter for low-complexity sequences was turned off, and the option `-task blastn-short` applied the optimised algorithm for short sequences. For each RRE, the mRNA hit with the lowest E-value was chosen, and with multiple mRNAs with lowest E-values, the longest transcript, was chosen. This procedure yielded 14,317 alignments with 5,843 mRNAs; only exact matches between sequenced RRE and mRNA hit were considered and thus for the remaining RREs only partial or no hits were identified.

For the remaining 14,317 RREs, we had to identify which of the top-expressed miRNA was most likely integrated into the bound Argonaute (if any). The supplementary material (Table S5) of [122] contains the expression profile of miRNAs in the same HEK293 cells that were used for the AGO1–4-`PAR-CLIP` experiment. Since an old `miRBase` [112,176] release was used,

[1] The BED format encodes genomic positions of annotations that can be viewed in tracks of genome browsers, such as the UCSC genome browser, `http://www.genome.ucsc.edu/FAQ/FAQformat.html`.

[2] BEDTools available from http://code.google.com/p/bedtools/

[3] `ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/refMrna.fa.gz`

mature miRNA IDs were updated via sequence comparisons to the `miRBase` release 19[1]. To reduce sequence redundancies when learning models of miRNA-MRE interactions, mature miRNAs were summarised into seed families when they shared identical seed sequences between positions 2–8. The miRNA with the highest observed expression level was chosen as the representative of each seed family. Finally, we used `IntaRNA` with restrictive parameter constraints to predict the exact MREs of the selected miRNAs within the 14,317 RREs. For this work, an unofficial version of `IntaRNA`, version 1.2.6 was applied so that when calculating miRNA-MRE interactions, there is a parameter that can restrict seed interaction within the target RNA as well as within the miRNA (which is not possible in the current official version). Thus, we could utilise the results from the previously described seed scanning to set the seed interaction with the target mRNA. Only seed positions that were within positions 20–30 of the AGO1–4-`PAR-CLIP` RREs were considered because this region was previously shown to be enriched in complementary seed matches to highly expressed miRNA [122], and this adds a further layer of precision to the data. To reflect the previous seed definitions, `IntaRNA` was set to a seed interaction of at least six base pairs including a maximum of two unpaired nucleotides. The calculation of accessibility was disabled, since it increases computational speed and enough evidence of the interaction site is given by the RREs detected by the AGO1–4-`PAR-CLIP` experiment. In the case of several `IntaRNA` hits, only the interaction with the minimum free energy was selected. `IntaRNA` interactions were subsequently annotated and filtered to belong to one of the four seed types defined in the beginning of this section. Finally, the predicted `IntaRNA` interactions were filtered further to increase the quality of the data according to the following criteria:

- The minimum free energy of the hybrid structure is <-4 kcal/mol.
- A maximum bulge size in the hybrid of 12 nt.
- Compensatory base pairing with at least four base pairs between positions 12–17. of the miRNA or pairing with positions 18–19 was required when only six base pairs existed in the seed region and at least three of these were *GU* base pairs; locations of compensatory base pairing for loose seed types was were taken from [166].

Curation of non-functional miRNA-MRE interactions

Since non-functional interactions between miRNA and MRE that do not affect target regulation are not published, such data has to be carefully generated. To select endogenous RNA sequences not targeted by miRNAs, we have assumed that those sequence regions of mRNAs, expressed in `CLIP-seq` experiments but where there is no evidence of cross-linking with Argonaute proteins, are not bound by miRNAs. Next, by masking the regions identified by `CLIP-seq` experiments, we searched for assumed non-functional interactions in the rest of the mRNA sequences.

[1] Downloaded April 2013 from `ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gx`

Collection of other CLIP-seq data. To have a large coverage of experimental evidence of miRNA-Argonaute binding, we collected all available CLIP-seq data from experiments performed on HEK293 cells. In addition to the AGO1–4 PAR-CLIP dataset from [122], we gathered data from three additional AGO2-CLIP-seq datasets from the Gene Expression Omnibus (GEO accession number is GSE28865) that were published in [172]. Of the two replicates performed, replicate A datasets were chosen due to their more consistent results [172]: these included GEO accessions GSM714642, GSM714644, and GSM714646. The three datasets contained 54,905, 91,362, and 44,497 40-nt-long sequences, respectively, with supporting evidence that they were bound by AGO2. The more recently developed protocol that ligates the miRNA sequence to the identified target sequence as well as cross-linking with Argonaute is CLASH [135]. The Supplementary Table S1 from [135] contains 18,514 miRNA-MRE interactions, however, these sequence vary in length from 18 to 119 nt, with the majority between 43–49 nt. As was previously done for the AGO1–4 PAR-CLIP dataset, exact locations on mRNA transcripts were identified by applying BLAST with an E-value cutoff of 0.0001 (and 0.001 for the CLASH dataset to capture the very short sequences). For a greater sensitivity, partial alignments were also considered for all CLIP-seq datasets.

Calculation of non-functional interactions. The aforementioned seed scanning was applied to all mRNAs for which an Argonaute-RRE was detected and all seed matches overlapping with these RREs from the five CLIP-seq datasets were ignored and hybrid interactions were calculated using IntaRNA with the same seed constraints as before. In order to select a negative set that closely resembles functional interactions, we selected interactions using the same filters applied to the functional interactions. To balance both datasets, the same number of seed types were added to the non-functional set that existed in the functional set with the same number of interactions per miRNA. A seed type was defined by the number of base pairs, specially regarding the number of *GU* base pairs, within the seed interaction: for example the seed type 6-2 would have 6 base-pairs within the seed interaction, two of which were *GU* base pairs.

D.4.3 RNA-binding-protein–occupancy profiles

The data produced for [12], downloaded from the Gene Expression Omnibus (GEO GSE38355), includes 4,740,558 nucleotide positions with evidence of cross-linking to any (unknown) RBP. This evidence is given by at least two of the characteristic T to C mutations that occur during the cDNA replication of cross-linked sites using the PAR-CLIP protocol [12, 122]. The genomic coordinates are given according to the older hg18 human genome assembly, and thus, the coordinates had to be mapped to the same mRNA transcripts used for the miRNA interaction datasets from the more recent hg19 genome assembly. This was done using the liftOver executable[1] and the respective conversion file from UCSC[2].

[1] liftOver downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/.

[2] Downloaded conversion file hg18ToHg19.over.chain from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/.

D.5 Part VI

D.5.1 GFP intensities measuring amiRNA efficacies for PIN1 in *A. thaliana*

GFP intensities were measured per protoplast in each sample separately. In Figure D.22, the raw intensity values are plotted. Since in each sample, we have multiple protoplasts where no GFP fluorescence was detected, we removed 50 % of the lowest values. The mean, median and standard devations are given for the top 50 % in each sample in table below. The controls are given by GFP-7 where GFP is fully expressed, a sample without any regulatory construct (mock/319a), and the wild type (wt) using the native ath-MIR319a (miRBase accession MI0000544). All amiRNAs P1–P62 and their complementary parts are inserted to the MIR319a precursor to measure their repression efficiencies. The median of the top 50 % corresponds to the 75 percentile of each sample. These values are normalised to lie between 0 (functional) to 1 (non-functional) using the wt and mock/319a samples as examples of functional and non-functional GFP intensities. The normalised values are given in the 'efficacy' column and are used for all analyses of function.

amiRNA	mean	median	standard deviation	efficacy
GFP-7	4.53	4.35	1.10	0.14
mock/319a	21.38	18.27	10.74	1.00
P01	3.53	3.32	1.02	0.08
P02	5.03	4.81	1.25	0.17
P03	9.92	8.57	3.09	0.40
P04	6.89	6.23	2.23	0.26
P05	18.03	16.80	8.25	0.91
P06	14.99	14.12	4.72	0.74
P07	12.29	11.15	4.13	0.56
P08	9.91	9.16	3.27	0.44
P09	8.26	8.06	2.46	0.37
P10	5.30	4.80	2.09	0.17
P11	9.57	8.48	3.12	0.40
P12	6.69	6.27	1.83	0.26
P13	8.20	7.66	2.11	0.35
P14	11.34	9.86	3.98	0.48
P16	6.68	5.99	1.62	0.24
P17	8.39	7.88	2.10	0.36
P18	10.13	9.13	3.19	0.44
P19	5.28	5.14	1.69	0.19
P20	7.00	6.53	2.53	0.28
P21	9.74	8.88	2.73	0.42
P22	6.45	5.78	2.05	0.23
P23	8.85	8.39	2.15	0.39
P24	10.78	9.29	3.62	0.45
P25	9.45	8.68	3.31	0.41
P26	10.05	8.88	2.92	0.42
P27	7.75	7.30	2.07	0.32
P28	17.11	16.19	5.70	0.87
P29	7.46	7.26	1.84	0.32

P30	10.19	9.12	3.83	0.44
P31	10.14	9.38	3.11	0.45
P32	7.25	6.77	1.97	0.29
P33	5.24	4.38	1.93	0.14
P34	18.90	15.54	11.31	0.83
P35	14.03	12.17	5.92	0.62
P36	12.22	10.78	4.29	0.54
P37	9.08	8.67	2.90	0.41
P38	6.11	5.75	1.56	0.23
P39	12.77	12.08	4.46	0.62
P40	10.83	9.10	6.19	0.43
P41	12.07	10.32	4.48	0.51
P42	6.59	6.59	1.41	0.28
P43	6.47	5.80	1.88	0.23
P44	7.23	6.42	2.03	0.27
P45	10.15	9.21	3.16	0.44
P46	8.00	7.38	2.43	0.33
P47	11.91	11.25	3.35	0.57
P48	12.79	11.40	3.78	0.58
P49	20.35	17.77	8.69	0.97
P50	9.40	8.39	2.84	0.39
P51	10.65	10.34	2.10	0.51
P52	4.64	4.11	1.94	0.13
P53	10.92	10.90	3.18	0.54
P54	13.93	12.30	4.51	0.63
P55	9.55	8.81	2.71	0.42
P56	17.87	15.31	7.10	0.82
P57	19.15	17.80	6.66	0.97
P58	6.52	5.64	2.11	0.22
P59	12.84	11.84	4.60	0.60
P60	10.72	10.60	3.00	0.53
P61	9.99	9.92	2.53	0.48
P62	12.72	12.59	4.13	0.65
wt	2.09	2.06	0.61	0.00

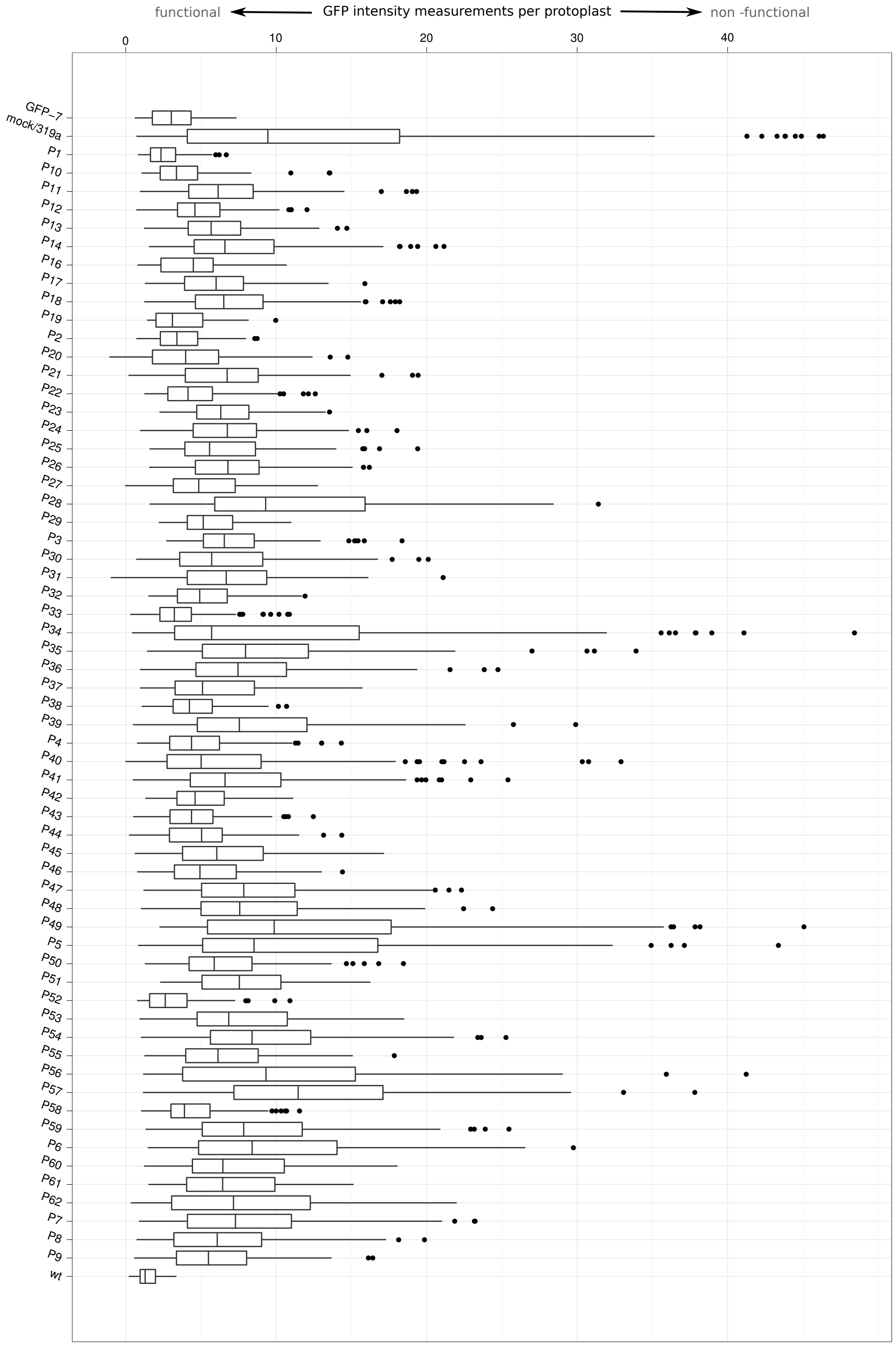


Figure D.22. Summary of raw GFP-intensity measurements for single protoplasts in the analysis of amiRNA efficacy. The y-axis shows different samples testing the efficacy of 62 different amiRNAs and the controls. What we observe is that for every sample, many protoplasts exist for which zero or very low GFP intensities were measured—irrespective of whether the amiRNA was functional (low intensities expected) or not functional (high intensities expected).

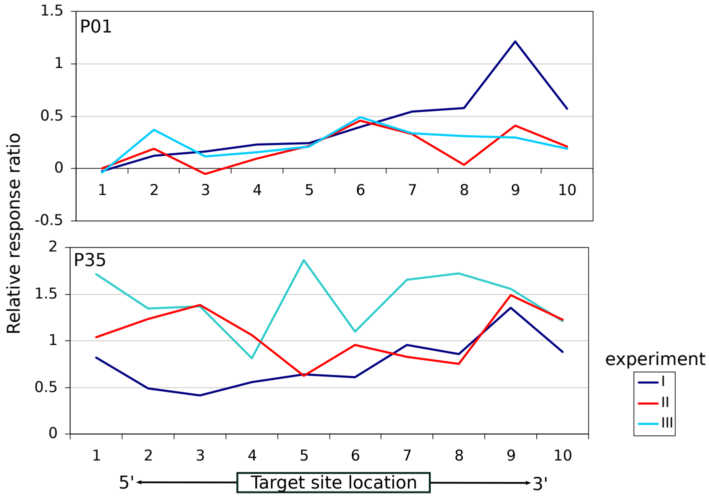


Figure D.23. Three independent replicate experiments for the ATGR2 gene in *A. thaliana*. Target sites of selected amiRNA (P01 and P35) were introduced into ten positions (x-axis) spread throughout the coding sequence of ATGR2 (as described in Section 10.2). Repression efficiencies are plotted on the y-axis as the relative response ratio, defined in Section 10.1.1. Experiments II and III correlate significantly for P01 with a Pearson's correlation coefficient of 0.75 (p=0.01); other combinations do not display significant correlations. For subsequent analyses, we chose the experiment with the lowest standard deviation in relative response ratios: experiment II for P01 and III for P35.

List of my publications

- [P1] Alkhnbashi OS, Costa F, Shah SA, Garrett RA, Saunders SJ, and Backofen R: **CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci**. *Bioinformatics* 2014, 30(17):i489–i496. In the proceedings of the 13th European Conference on Computational Biology (ECCB) 2014.
- [P2] Brendel J, Stoll B, Lange SJ, Sharma K, Lenz C, Stachler AE, Maier LK, Richter H, Nickel L, Schmitz RA, *et al.*: **A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of crRNAs in Haloferax volcanii**. *Journal of Biological Chemistry* 2014, 289(10):7164–77.
- [P3] Lange SJ, Alkhnbashi OS, Rose D, Will S, and Backofen R: **CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems**. *Nucleic Acids Res* 2013, 41(17):8034–44. SJL, OSA and DR contributed equally to this work.
- [P4] Lange SJ, Maticzka D, Möhl M, Gagnon JN, Brown CM, and Backofen R: **Global or local? Predicting secondary structure and accessibility in mRNAs**. *Nucleic Acids Res* 2012, 40(12):5215–26. SJL and DM contributed equally to this work.
- [P5] Maier LK, Lange SJ, Stoll B, Haas KA, Fischer S, Fischer E, Duchardt-Ferner E, Wohnert J, Backofen R, and Marchfelder A: **Essential requirements for the detection and degradation of invaders by the Haloferax volcanii CRISPR/Cas system I-B**. *RNA Biol* 2013, 10(5).
- [P6] Maticzka D, Lange SJ, Costa F, and Backofen R: **GraphProt: modeling binding preferences of RNA-binding proteins**. *Genome Biol* 2014, 15(1):R17.
- [P7] Nickel L, Weidenbach K, Jager D, Backofen R, Lange SJ, Heidrich N, and Schmitz RA: **Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III**. *RNA Biol* 2013, 10(5):779–791.
- [P8] Reimann V, Saunders SJ, Scholz I, Alkhnbashi OS, Hein S, Backofen R, and Hess WR: **Structural constraints and enzymatic promiscuity in the cas6-dependent generation of crRNAs in cyanobacteria**. *RNA* 2014. Submitted.
- [P9] Richter H, Lange SJ, Backofen R, and Randau L: **SF CRISPR: Comparative analysis of Cas6b processing and CRISPR RNA stability**. *RNA Biol* 2013, 10(5).
- [P10] Scholz I, Lange SJ, Hein S, Hess WR, and Backofen R: **CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein**. *PLoS One* 2013, 8(2):e56 470. IS and SJL contributed equally to this work.
- [P11] Stoll B, Maier LK, Lange SJ, Brendel J, Fischer S, Backofen R, and Marchfelder A: **Requirements for a successful defence reaction by the CRISPR-Cas subtype I-B system**. *Biochem Soc Trans* 2013, 41(6):1444–8.

Bibliography

[1] Agius C, Eamens AL, Millar AA, Watson JM, and Wang MB: **RNA silencing and antiviral defense in plants**. *Methods Mol Biol* 2012, 894:17–38.

[2] Al-Attar S, Westra ER, van der Oost J, and Brouns SJJ: **Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes**. *Biol Chem* 2011, 392(4):277–89.

[3] Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, and Hatzigeorgiou AG: **Lost in translation: an assessment and perspective for computational microRNA target identification**. *Bioinformatics* 2009, 25(23):3049–55.

[4] Altman S: **Biosynthesis of transfer RNA in Escherichia coli**. *Cell* 1975, 4(1):21–9.

[5] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, 215(3):403–10.

[6] Ameres SL and Zamore PD: **Diversifying microRNA sequence and function**. *Nat Rev Mol Cell Biol* 2013, 14(8):475–88.

[7] Andronescu M, Bereg V, Hoos HH, and Condon A: **RNA STRAND: the RNA secondary structure and statistical analysis database**. *BMC Bioinformatics* 2008, 9:340.

[8] Arslan Z, Hermanns V, Wurm R, Wagner R, and Pul U: **Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system**. *Nucleic Acids Res* 2014, 42(12):7884–93.

[9] Ascano M, Hafner M, Cekan P, Gerstberger S, and Tuschl T: **Identification of RNA-protein interaction networks using PAR-CLIP**. *Wiley Interdiscip Rev RNA* 2012, 3(2):159–77.

[10] Axtell MJ, Westholm JO, and Lai EC: **Vive la difference: biogenesis and evolution of microRNAs in plants and animals**. *Genome Biol* 2011, 12(4):221.

[11] Baek D, Villen J, Shin C, Camargo FD, Gygi SP, and Bartel DP: **The impact of microRNAs on protein output**. *Nature* 2008, 455(7209):64–71.

[12] Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, *et al.*: **The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts**. *Mol Cell* 2012, 46(5):674–90.

[13] Bandyopadhyay S and Mitra R: **TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples**. *Bioinformatics* 2009, 25(20):2625–31.

[14] Barreau C, Paillard L, and Osborne HB: **AU-rich elements and associated factors: are there unifying principles?** *Nucleic Acids Res* 2005, 33(22):7138–50.

[15] Becker C: *RNAi-mediated gene silencing by small non-coding RNAs in protoplasts of Arabidopsis thaliana.* Ph.D. thesis, Albert-Ludwigs-University Freiburg, 2010.

[16] Bernhart SH, Hofacker IL, and Stadler PF: **Local RNA base pairing probabilities in large sequences**. *Bioinformatics* 2006, 22(5):614–5.

[17] Bernhart SH, Hofacker IL, Will S, Gruber AR, and Stadler PF: **RNAalifold: improved consensus structure prediction for RNA alignments**. *BMC Bioinformatics* 2008, 9:474.

[18] Bernhart SH, Mückstein U, and Hofacker IL: **RNA Accessibility in cubic time**. *Algorithms Mol Biol* 2011, 6(1):3.

[19] Bernhart SH, Tafer H, Muckstein U, Flamm C, Stadler PF, and Hofacker IL: **Partition function and base pairing probabilities of RNA heterodimers**. *Algorithms Mol Biol* 2006, 1(1):3.

[20] Betel D, Koppal A, Agius P, Sander C, and Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome Biol* 2010, 11(8):R90.

[21] Bhattacharyya SN, Habermacher R, Martine U, Closs EI, and Filipowicz W: **Relief of microRNA-mediated translational repression in human cells subjected to stress**. *Cell* 2006, 125(6):1111–24.

[22] Biswas A, Fineran P, and Brown C: **Accurate computational prediction of the transcribed strand of CRISPR noncoding RNAs**. *Bioinformatics* 2014.

[23] Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, and Hugenholtz P: **CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats**. *BMC Bioinformatics* 2007, 8:209.

[24] Blencowe BJ, Ahmad S, and Lee LJ: **Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes**. *Genes Dev* 2009, 23(12):1379–86.

[25] Bokov K and Steinberg SV: **A hierarchical model for evolution of 23S ribosomal RNA**. *Nature* 2009, 457(7232):977–80.

[26] Bolognani F and Perrone-Bizzozero NI: **RNA-protein interactions and control of mRNA stability in neurons**. *J Neurosci Res* 2008, 86(3):481–9.

[27] Bon M and Orland H: **TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots**. *Nucleic Acids Res* 2011, 39(14):e93.

[28] Breaker RR: **Complex riboswitches**. *Science* 2008, 319(5871):1795–7.

[29] Breaker RR: **Riboswitches and the RNA world**. *Cold Spring Harb Perspect Biol* 2012, 4(2):a003 566.

[30] Broderick JA, Salomon WE, Ryder SP, Aronin N, and Zamore PD: **Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing**. *RNA* 2011, 17(10):1858–69.

[31] Brodt A, Lurie-Weinberger MN, and Gophna U: **CRISPR loci reveal networks of gene exchange in archaea**. *Biol Direct* 2011, 6(1):65.

[32] Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, and van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes**. *Science* 2008, 321(5891):960–4.

[33] Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, and Bateman A: **Rfam 11.0: 10 years of RNA families**. *Nucleic Acids Res* 2013, 41(Database issue):D226–32.

[34] Busch A, Richter AS, and Backofen R: **IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions**. *Bioinformatics* 2008, 24(24):2849–56.

[35] Cao S and Chen SJ: **Predicting kissing interactions in microRNA-target complex and assessment of microRNA activity**. *Nucleic Acids Res* 2012, 40(10):4681–90.

[36] Carte J, Pfister NT, Compton MM, Terns RM, and Terns MP: **Binding and cleavage of CRISPR RNA by Cas6**. *RNA* 2010, 16(11):2181–8.

[37] Carvalho LE and Lawrence CE: **Centroid estimation in discrete high-dimensional spaces with applications in biology**. *Proc Natl Acad Sci USA* 2008, 105(9):3209–14.

[38] Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, *et al.*: **Insights into RNA biology from an atlas of mammalian mRNA-binding proteins**. *Cell* 2012, 149(6):1393–406.

[39] Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, and Bozzoni I: **A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA**. *Cell* 2011, 147(2):358–69.

[40] Chan PP and Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence**. *Nucleic Acids Res* 2009, 37(Database issue):D93–7.

[41] Chandra V, Girijadevi R, Nair AS, Pillai SS, and Pillai RM: **MTar: a computational microRNA target prediction architecture for human transcriptome**. *BMC Bioinformatics* 2010, 11 Suppl 1:S2.

[42] Chang SH, Lu YC, Li X, Hsieh WY, Xiong Y, Ghosh M, Evans T, Elemento O, and Hla T: **Antagonistic function of the RNA-binding protein HuR and miR-200b in post-transcriptional regulation of vascular endothelial growth factor-A expression and angiogenesis**. *Journal of Biological Chemistry* 2013, 288(7):4908–21.

[43] Charikar M, Guruswami V, Kumar R, Rajagopalan S, and Sahai A: **Combinatorial feature selection problems**. In IEEE (ed.), **41st Annual Symposium on Foundations of Computer Science: proceedings: 12–14 November, 2000, Redondo Beach, California** (IEEE Computer Society Press, 1109 Spring Street, Suite 300, Silver Spring, MD 20910, USA2000), ISBN 0-7695-0850-2, 0-7695-0851-0 (case), 0-7695-0852-9 (microfiche) pp. 631–640.

[44] Cheng L, Quek CYJ, Sun X, Bellingham SA, and Hill AF: **The detection of microRNA associated with Alzheimer's disease in biological fluids using next-generation sequencing technologies**. *Front Genet* 2013, 4:150.

[45] Chi SW, Hannon GJ, and Darnell RB: **An alternative mode of microRNA target recognition**. *Nat Struct Mol Biol* 2012, 19(3):321–7.

[46] Chi SW, Zang JB, Mele A, and Darnell RB: **Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps**. *Nature* 2009, 460(7254):479–86.

[47] Chitsaz H, Salari R, Sahinalp SC, and Backofen R: **A partition function algorithm for interacting nucleic acid strands**. *Bioinformatics* 2009, 25(12):i365–73.

[48] Chojnowski G, Walen T, and Bujnicki JM: **RNA Bricks–a database of RNA 3D motifs and their interactions**. *Nucleic Acids Res* 2014, 42(Database issue):D123–31.

[49] Chou CH, Lin FM, Chou MT, Hsu SD, Chang TH, Weng SL, Shrestha S, Hsiao CC, Hung JH, and Huang HD: **A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing**. *BMC Genomics* 2013, 14 Suppl 1:S2.

[50] Cock PJA, Fields CJ, Goto N, Heuer ML, and Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Res* 2010, 38(6):1767–71.

[51] Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, *et al.*: **The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy**. *Nucleic Acids Res* 2003, 31(1):442–3.

[52] Conrad NK: **The emerging role of triple helices in RNA biology**. *Wiley Interdiscip Rev RNA* 2014, 5(1):15–29.

[53] Consortium IHGS: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, 431(7011):931–45.

[54] Consortium TEP: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, 447(7146):799–816.

[55] Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, and Ohler U: **PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data**. *Genome Biol* 2011, 12(8):R79.

[56] Costa F and Grave KD: **Fast neighborhood subgraph pairwise distance kernel**. In **Proceedings of the 26 th International Conference on Machine Learning** (Omnipress2010) pp. 255–262.

[57] Crooks GE, Hon G, Chandonia JM, and Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, 14(6):1188–90.

[58] Darnell RB: **HITS-CLIP: panoramic views of protein-RNA regulation in living cells**. *Wiley Interdiscip Rev RNA* 2010, 1(2):266–86.

[59] Deigan KE, Li TW, Mathews DH, and Weeks KM: **Accurate SHAPE-directed RNA structure determination**. *Proc Natl Acad Sci USA* 2009, 106(1):97–102.

[60] Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, and Charpentier E: **CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III**. *Nature* 2011, 471(7340):602–7.

[61] Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, and Ribeca P: **Fast computation and applications of genome mappability**. *PLoS One* 2012, 7(1):e30 377.

[62] DeSantis TZ, Keller K, Karaoz U, Alekseyenko AV, Singh NNS, Brodie EL, Pei Z, Andersen GL, and Larsen N: **Simrank: Rapid and sensitive general-purpose k-mer search tool**. *BMC Ecol* 2011, 11:11.

[63] Deveson I, Li J, and Millar AA: **MicroRNAs with analogous target complementarities perform with highly variable efficacies in Arabidopsis**. *FEBS Lett* 2013, 587(22):3703–8.

[64] Diamond JM, Turner DH, and Mathews DH: **Thermodynamics of three-way multibranch loops in RNA**. *Biochemistry* 2001, 40(23):6971–81.

[65] Didiano D and Hobert O: **Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions**. *Nat Struct Mol Biol* 2006, 13(9):849–51.

[66] Ding L and Han M: **GW182 family proteins are crucial for microRNA-mediated gene silencing**. *Trends Cell Biol* 2007, 17(8):411–6.

[67] Ding Y, Chan CY, and Lawrence CE: **Sfold web server for statistical folding and rational design of nucleic acids**. *Nucleic Acids Res* 2004, 32(Web Server issue):W135–41.

[68] Ding Y, Chan CY, and Lawrence CE: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble**. *RNA* 2005, 11(8):1157–66.

[69] Ding Y, Chan CY, and Lawrence CE: **Clustering of RNA secondary structures with application to messenger RNAs**. *J Mol Biol* 2006, 359(3):554–71.

[70] Ding Y and Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction**. *Nucleic Acids Res* 2003, 31(24):7280–301.

[71] Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, and Assmann SM: **In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features**. *Nature* 2014, 505(7485):696–700.

[72] Do CB, Woods DA, and Batzoglou S: **CONTRAfold: RNA secondary structure prediction without physics-based models**. *Bioinformatics* 2006, 22(14):e90–8.

[73] Doose G and Metzler D: **Bayesian sampling of evolutionarily conserved RNA secondary structures with pseudoknots**. *Bioinformatics* 2012, 28(17):2242–8.

[74] Doshi KJ, Cannone JJ, Cobaugh CW, and Gutell RR: **Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction**. *BMC Bioinformatics* 2004, 5:105.

[75] Eddy SR: **Accelerated Profile HMM Searches**. *PLoS Comput Biol* 2011, 7(10):e1002 195.

[76] Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, 5:113.

[77] Eiring AM, Harb JG, Neviani P, Garton C, Oaks JJ, Spizzo R, Liu S, Schwind S, Santhanam R, Hickey CJ, *et al.*: **mir-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts**. *Cell* 2010, 140(5):652–65.

[78] Elkayam E, Kuhn CD, Tocilj A, Haase AD, Greene EM, Hannon GJ, and Joshua-Tor L: **The Structure of Human Argonaute-2 in Complex with miR-20a**. *Cell* 2012, 150(1):100–10.

[79] Enright AJ, John B, Gaul U, Tuschl T, Sander C, and Marks DS: **MicroRNA targets in Drosophila**. *Genome Biol* 2003, 5(1):R1.

[80] Enright AJ, Van Dongen S, and Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res* 2002, 30(7):1575–84.

[81] Farazi TA, Hoell JI, Morozov P, and Tuschl T: **MicroRNAs in human cancer**. *Adv Exp Med Biol* 2013, 774:1–20.

[82] Fawcett T: **An introduction to roc analysis**. *Pattern Recognition Letters* 2006, 27(8):861 – 874. ROC Analysis in Pattern Recognition.

[83] Fedorov Y, Anderson EM, Birmingham A, Reynolds A, Karpilow J, Robinson K, Leake D, Marshall WS, and Khvorova A: **Off-target effects by siRNA can induce toxic phenotype**. *RNA* 2006, 12(7):1188–96.

[84] Feng DF and Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees**. *J Mol Evol* 1987, 25(4):351–60.

[85] Fields DS and Gutell RR: **An analysis of large rRNA sequences folded by a thermodynamic method**. *Fold Des* 1996, 1(6):419–30.

[86] Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2008, 36(Database issue):D281–8.

[87] Fjose A and Drivenes O: **RNAi and microRNAs: from animal models to disease therapy**. *Birth Defects Res C Embryo Today* 2006, 78(2):150–71.

[88] Flamm C and Hofacker I: **Beyond energy minimization: approaches to the kinetic folding of RNA**. *Chemical Monthly* 2008, 139:447–457.

[89] Forch P, Puig O, Kedersha N, Martinez C, Granneman S, Seraphin B, Anderson P, and Valcarcel J: **The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing**. *Mol Cell* 2000, 6(5):1089–98.

[90] Frasconi P, Costa F, Raedt LD, and Grave KD: **klog: A language for logical and relational learning with kernels**. *CoRR* 2012, abs/1205.3981.

[91] Freeberg MA, Han T, Moresco JJ, Kong A, Yang YC, Lu ZJ, Yates JR, and Kim JK: **Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae**. *Genome Biol* 2013, 14(2):R13.

[92] Friedman RC, Farh KKH, Burge CB, and Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, 19(1):92–105.

[93] Gaidatzis D, van Nimwegen E, Hausser J, and Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathway analysis**. *BMC Bioinformatics* 2007, 8:69.

[94] Garcia DM, Baek D, Shin C, Bell GW, Grimson A, and Bartel DP: **Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs**. *Nat Struct Mol Biol* 2011, 18(10):1139–46.

[95] Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, *et al.*: **Rfam: Wikipedia, clans and the ”decimal” release**. *Nucleic Acids Res* 2011, 39(Database issue):D141–5.

[96] Gardner PP, Wilm A, and Washietl S: **A benchmark of multiple sequence alignment programs upon structural RNAs**. *Nucleic Acids Res* 2005, 33(8):2433–9.

[97] Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, and Moineau S: **The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA**. *Nature* 2010, 468(7320):67–71.

[98] Garrett RA, Vestergaard G, and Shah SA: **Archaeal CRISPR-based immune systems: exchangeable functional modules**. *Trends Microbiol* 2011, 19(11):549–56.

[99] Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, and MacMillan AM: **Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases**. *RNA* 2012, 18(11):2020–8.

[100] Gatignol A, Buckler C, and Jeang KT: **Relatedness of an RNA-binding motif in human immunodeficiency virus type 1 TAR RNA-binding protein TRBP to human P1/dsI kinase and Drosophila staufen**. *Mol Cell Biol* 1993, 13(4):2193–202.

[101] Geisler S and Coller J: **RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts**. *Nat Rev Mol Cell Biol* 2013, 14(11):699–712.

[102] German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K, German R, *et al.*: **Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends**. *Nat Biotechnol* 2008, 26(8):941–6.

[103] Gesner EM, Schellenberg MJ, Garside EL, George MM, and Macmillan AM: **Recognition and maturation of effector RNAs in a CRISPR interference pathway**. *Nat Struct Mol Biol* 2011, 18(6):688–92.

[104] Gibcus JH and Dekker J: **The context of gene expression regulation**. *F1000 Biol Rep* 2012, 4:8.

[105] Giege R, Juhling F, Putz J, Stadler P, Sauter C, and Florentz C: **Structure of transfer RNAs: similarity and variability**. *Wiley Interdiscip Rev RNA* 2012, 3(1):37–61.

[106] Glisovic T, Bachorik JL, Yong J, and Dreyfuss G: **RNA-binding proteins and post-transcriptional gene regulation**. *FEBS Lett* 2008, 582(14):1977–86.

[107] Gogleva AA, Gelfand MS, and Artamonova II: **Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs**. *BMC Genomics* 2014, 15:202.

[108] Gorodkin J and Hofacker IL: **From structure prediction to genomic screens for novel non-coding RNAs**. *PLoS Comput Biol* 2011, 7(8):e1002 100.

[109] Gredell JA, Berger AK, and Walton SP: **Impact of target mRNA structure on siRNA silencing efficiency: A large-scale study**. *Biotechnol Bioeng* 2008, 100(4):744–55.

[110] Gregory SG, Barlow KF, McLay KE, Kaul R, Swarbreck D, Dunham A, Scott CE, Howe KL, Woodfine K, Spencer CCA, *et al.*: **The DNA sequence and biological annotation of human chromosome 1**. *Nature* 2006, 441(7091):315–21.

[111] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res* 2005, 33 Database Issue:D121–4.

[112] Griffiths-Jones S, Saini HK, van Dongen S, and Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Res* 2008, 36(Database issue):D154–8.

[113] Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, and Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. *Mol Cell* 2007, 27(1):91–105.

[114] Grissa I, Vergnaud G, and Pourcel C: **The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats**. *BMC Bioinformatics* 2007, 8:172.

[115] Grissa I, Vergnaud G, and Pourcel C: **CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats**. *NAR* 2007, 35(Web Server issue):W52–7.

[116] Gronau I and Moran S: **Optimal implementations of upgma and other common clustering algorithms**. *Inf Process Lett* 2007, 104(6):205–210.

[117] Gruber AR, Bernhart SH, Hofacker IL, and Washietl S: **Strategies for measuring evolutionary conservation of RNA secondary structures**. *BMC Bioinformatics* 2008, 9:122.

[118] Gruber AR, Findeiss S, Washietl S, Hofacker IL, and Stadler PF: **RNAZ 2.0: IMPROVED NON-CODING RNA DETECTION**. In **PSB10**, vol. 15 (2010) pp. 69–79.

[119] Gruber AR, Lorenz R, Bernhart SH, Neubock R, and Hofacker IL: **The Vienna RNA websuite**. *Nucleic Acids Res* 2008, 36(Web Server issue):W70–4.

[120] Gustafson AM, Allen E, Givan S, Smith D, Carrington JC, and Kasschau KD: **ASRP: the Arabidopsis Small RNA Project Database**. *Nucleic Acids Res* 2005, 33(Database issue):D637–40.

[121] Ha I, Wightman B, and Ruvkun G: **A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation**. *Genes Dev* 1996, 10(23):3041–50.

[122] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano MJ, Jungkamp AC, Munschauer M, *et al.*: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP**. *Cell* 2010, 141(1):129–41.

[123] Haft DH, Selengut J, Mongodin EF, and Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes**. *PLoS Comput Biol* 2005, 1(6):e60.

[124] Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, and Beck E: **TIGRFAMs and Genome Properties in 2013**. *Nucleic Acids Res* 2013, 41(Database issue):D387–95.

[125] Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, Resch AM, Glover CVCr, Graveley BR, Terns RM, *et al.*: **Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs**. *Mol Cell* 2012, 45(3):292–302.

[126] Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, and Terns MP: **RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex**. *Cell* 2009, 139(5):945–56.

[127] Han K and Nepal C: **PRI-Modeler: extracting RNA structural elements from PDB files of protein-RNA complexes**. *FEBS Lett* 2007, 581(9):1881–90.

[128] Hannon GJ: **RNA interference**. *Nature* 2002, 418(6894):244–51.

[129] Hatoum-Aslan A, Maniv I, and Marraffini LA: **Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site**. *Proc Natl Acad Sci USA* 2011, 108(52):21 218–22.

[130] Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, and Doudna JA: **Sequence- and structure-specific RNA processing by a CRISPR endonuclease**. *Science* 2010, 329(5997):1355–8.

[131] Haurwitz RE, Sternberg SH, and Doudna JA: **Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA**. *EMBO J* 2012, 31(12):2824–32.

[132] Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, and Zavolan M: **Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets**. *Genome Res* 2009, 19(11):2009–20.

[133] Havgaard JH, Lyngso RB, Stormo GD, and Gorodkin J: **Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%**. *Bioinformatics* 2005, 21(9):1815–24.

[134] Havgaard JH, Torarinsson E, and Gorodkin J: **Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix**. *PLoS Comput Biol* 2007, 3(10):1896–908.

[135] Helwak A, Kudla G, Dudnakova T, and Tollervey D: **Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding**. *Cell* 2013, 153(3):654–65.

[136] Heyne S, Costa F, Rose D, and Backofen R: **GraphClust: alignment-free structural clustering of local RNA secondary structures**. *Bioinformatics* 2012, 28(12):i224–i232.

[137] Hofacker IL, Bernhart SH, and Stadler PF: **Alignment of RNA base pairing probability matrices**. *Bioinformatics* 2004, 20(14):2222–7.

[138] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, and Schuster P: **Fast folding and comparison of RNA secondary structures**. *Monatshefte Chemie* 1994, 125:167–188.

[139] Hofacker IL, Priwitzer B, and Stadler PF: **Prediction of locally stable RNA secondary structures for genome-wide surveys**. *Bioinformatics* 2004, 20(2):186–190.

[140] Hofacker IL and Stadler PF: **Memory efficient folding algorithms for circular RNA secondary structures**. *Bioinformatics* 2006, 22(10):1172–6.

[141] Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, and Hackermuller J: **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. *PLoS Comput Biol* 2009, 5(9):e1000 502.

[142] Honer zu Siederdissen C and Hofacker IL: **Discriminatory power of RNA family models**. *Bioinformatics* 2010, 26(18):i453–9.

[143] Hong X, Hammell M, Ambros V, and Cohen SM: **Immunopurification of Ago1 miRNPs selects for a distinct class of microRNA targets**. *Proc Natl Acad Sci USA* 2009, 106(35):15 085–90.

[144] Horvath P, Coute-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, and Barrangou R: **Comparative analysis of CRISPR loci in lactic acid bacteria genomes**. *Int J Food Microbiol* 2009, 131(1):62–70.

[145] Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, *et al.*: **DNA targeting specificity of RNA-guided Cas9 nucleases**. *Nat Biotechnol* 2013, 31(9):827–32.

[146] Huang FWD, Qin J, Reidys CM, and Stadler PF: **Partition function and base pairing probabilities for RNA-RNA interaction prediction**. *Bioinformatics* 2009, 25(20):2646–54.

[147] Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, Tajnik M, Konig J, and Ule J: **iCLIP: Protein-RNA interactions at nucleotide resolution**. *Methods* 2013.

[148] Incarnato D, Neri F, Diamanti D, and Oliviero S: **MREdictor: a two-step dynamic interaction model that accounts for mRNA accessibility and Pumilio binding accurately predicts microRNA targets**. *Nucleic Acids Res* 2013, 41(18):8421–33.

[149] Jacobs GH, Chen A, Stevens SG, Stockwell PA, Black MA, Tate WP, and Brown CM: **Transterm: a database to aid the analysis of regulatory sequences in mRNAs**. *Nucleic Acids Res* 2009, 37(Database issue):D72–6.

[150] Jacobsen A, Wen J, Marks DS, and Krogh A: **Signatures of RNA binding proteins globally coupled to effective microRNA target sites**. *Genome Res* 2010, 20(8):1010–9.

[151] Jansen R, Embden JDAv, Gaastra W, and Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes**. *Mol Microbiol* 2002, 43(6):1565–75.

[152] Jenkins RH, Bennagi R, Martin J, Phillips AO, Redman JE, and Fraser DJ: **A conserved stem loop motif in the 5'untranslated region regulates transforming growth factor-beta(1) translation**. *PLoS One* 2010, 5(8):e12 283.

[153] John B, Enright AJ, Aravin A, Tuschl T, Sander C, and Marks DS: **Human MicroRNA targets**. *PLoS Biol* 2004, 2(11):e363.

[154] Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, *et al.*: **Structural basis for CRISPR RNA-guided DNA recognition by Cascade**. *Nat Struct Mol Biol* 2011, 18(5):529–36.

[155] Juranek S, Eban T, Altuvia Y, Brown M, Morozov P, Tuschl T, and Margalit H: **A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs**. *RNA* 2012, 18(4):783–94.

[156] Jurica MS, Sousa D, Moore MJ, and Grigorieff N: **Three-dimensional structure of C complex spliceosomes by electron microscopy**. *Nat Struct Mol Biol* 2004, 11(3):265–269.

[157] Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, and Gorodkin J: **Structural profiles of human miRNA families from pairwise clustering**. *Bioinformatics* 2009, 25(3):291–4.

[158] Kaneko T, Nakamura Y, Sasamoto S, Watanabe A, Kohara M, Matsumoto M, Shimpo S, Yamada M, and Tabata S: **Structural analysis of four large plasmids harboring in a unicellular cyanobacterium, Synechocystis sp. PCC 6803**. *DNA Res* 2003, 10(5):221–8.

[159] Katoh K, Misawa K, Kuma Ki, and Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, 30(14):3059–66.

[160] Katoh K and Toh H: **Recent developments in the MAFFT multiple sequence alignment program**. *Brief Bioinform* 2008, 9(4):286–98.

[161] Kazan H, Ray D, Chan ET, Hughes TR, and Morris Q: **RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins**. *PLoS Comput Biol* 2010, 6:e1000 832.

[162] Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JAF, Elkon R, and Agami R: **A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility**. *Nat Cell Biol* 2010, 12(10):1014–20.

[163] Kel A, Voss N, Jauregui R, Kel-Margoulis O, and Wingender E: **Beyond microarrays: Finding key transcription factors controlling signal transduction pathways**. *BMC Bioinformatics* 2006, 7 Suppl 2:S13.

[164] Kertesz M, Iovino N, Unnerstall U, Gaul U, and Segal E: **The role of site accessibility in microRNA target recognition**. *Nat Genet* 2007, 39(10):1278–84.

[165] Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, and Segal E: **Genome-wide measurement of RNA secondary structure in yeast**. *Nature* 2010, 467(7311):103–7.

[166] Khorshid M, Hausser J, Zavolan M, and van Nimwegen E: **A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets**. *Nat Methods* 2013, 10(3):253–5.

[167] Khraiwesh B, Ossowski S, Weigel D, Reski R, and Frank W: **Specific gene silencing by artificial MicroRNAs in Physcomitrella patens: an alternative to targeted gene knockouts**. *Plant Physiol* 2008, 148(2):684–93.

[168] Kim HS, Headey SJ, Yoga YMK, Scanlon MJ, Gorospe M, Wilce MCJ, and Wilce JA: **Distinct binding properties of TIAR RRMs and linker region**. *RNA Biol* 2013, 10(4):579–89.

[169] Kircher M and Kelso J: **High-throughput DNA sequencing–concepts and limitations**. *Bioessays* 2010, 32(6):524–36.

[170] Kiryu H, Kin T, and Asai K: **Rfold: an exact algorithm for computing local base pairing probabilities**. *Bioinformatics* 2008, 24(3):367–73.

[171] Kiryu H, Terai G, Imamura O, Yoneyama H, Suzuki K, and Asai K: **A detailed investigation of accessibilities around target sites of siRNAs and miRNAs**. *Bioinformatics* 2011, 27(13):1788–97.

[172] Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, and Zavolan M: **A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins**. *Nat Methods* 2011, 8(7):559–64.

[173] Knudsen B and Hein J: **Pfold: RNA secondary structure prediction using stochastic context-free grammars**. *Nucleic Acids Res* 2003, 31(13):3423–8.

[174] Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, and Ule J: **iCLIP–transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution**. *J Vis Exp* 2011, (50).

[175] Konings DA and Gutell RR: **A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs**. *RNA* 1995, 1(6):559–74.

[176] Kozomara A and Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic Acids Res* 2014, 42(Database issue):D68–73.

[177] Krishnakumar R and Blelloch RH: **Epigenetics of cellular reprogramming**. *Curr Opin Genet Dev* 2013.

[178] Kruger J and Rehmsmeier M: **RNAhybrid: microRNA target prediction easy, fast and flexible**. *Nucleic Acids Res* 2006, 34(Web Server issue):W451–4.

[179] Krysan PJ, Young JC, and Sussman MR: **T-DNA as an insertional mutagen in Arabidopsis**. *Plant Cell* 1999, 11(12):2283–90.

[180] Kunin V, Sorek R, and Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats**. *Genome Biol* 2007, 8(4):R61.

[181] Kutter C and Svoboda P: **miRNA, siRNA, piRNA: Knowns of the unknown**. *RNA Biol* 2008, 5(4):181–8.

[182] Laing C and Schlick T: **Computational approaches to 3D modeling of RNA**. *J Phys Condens Matter* 2010, 22(28):283 101.

[183] Lall S, Grün D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, MacMenamin P, *et al.*: **A genome-wide map of conserved microRNA targets in *C. elegans***. *Curr Biol* 2006, 16(5):460–71.

[184] Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, *et al.*: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**. *Nucleic Acids Res* 2012, 40(Database issue):D1202–10.

[185] Lee JH, Kim H, Ko J, and Lee Y: **Interaction of C5 protein with RNA aptamers selected by SELEX**. *Nucleic Acids Res* 2002, 30(24):5360–8.

[186] Lee RC, Feinbaum RL, and Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, 75(5):843–54.

[187] Lekprasert P, Mayhew M, and Ohler U: **Assessing the Utility of Thermodynamic Features for microRNA Target Prediction under Relaxed Seed and No Conservation Requirements**. *PLoS One* 2011, 6(6):e20 622.

[188] Leontis NB and Westhof E: **Geometric nomenclature and classification of RNA base pairs**. *RNA* 2001, 7(4):499–512.

[189] Letunic I and Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy**. *Nucleic Acids Res* 2011, 39(Web Server issue):W475–8.

[190] Lewis BP, Burge CB, and Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, 120(1):15–20.

[191] Lewis BP, Shih Ih, Jones-Rhoades MW, Bartel DP, and Burge CB: **Prediction of mammalian microRNA targets**. *Cell* 2003, 115(7):787–98.

[192] Li J, Kim T, Nutiu R, Ray D, Hughes TR, and Zhang Z: **Identifying mRNA sequence elements for target recognition by human Argonaute proteins**. *Genome Res* 2014, 24(5):775–785.

[193] Li L, Huang D, Cheung MK, Nong W, Huang Q, and Kwan HS: **BSRD: a repository for bacterial small regulatory RNA**. *Nucleic Acids Res* 2013, 41(Database issue):D233–8.

[194] Li X, Quon G, Lipshitz HD, and Morris Q: **Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure**. *RNA* 2010, 16(6):1096–107.

[195] Ling H, Fabbri M, and Calin GA: **MicroRNAs and other non-coding RNAs as targets for anticancer drug development**. *Nat Rev Drug Discov* 2013, 12(11):847–65.

[196] Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB, Leslie CS, and Rudensky AY: **Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting**. *Mol Cell* 2012, 48(5):760–70.

[197] Lu ZJ, Gloor JW, and Mathews DH: **Improved RNA secondary structure prediction by maximizing expected pair accuracy**. *RNA* 2009, 15(10):1805–13.

[198] Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA, and Arkin AP: **Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)**. *Proc Natl Acad Sci USA* 2011, 108(27):11 063–8.

[199] Lunde BM, Moore C, and Varani G: **RNA-binding proteins: modular design for efficient function**. *Nat Rev Mol Cell Biol* 2007, 8(6):479–90.

[200] Ma L, Wei L, Wu F, Hu Z, Liu Z, and Yuan W: **Advances with microRNAs in Parkinson's disease research**. *Drug Des Devel Ther* 2013, 7:1103–13.

[201] Makarova KS, Aravind L, Wolf YI, and Koonin EV: **Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems**. *Biol Direct* 2011, 6:38.

[202] Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, *et al.*: **Evolution and classification of the CRISPR-Cas systems**. *Nat Rev Microbiol* 2011, 9(6):467–77.

[203] Mali P, Esvelt KM, and Church GM: **Cas9 as a versatile tool for engineering biology**. *Nat Methods* 2013, 10(10):957–63.

[204] Marin RM and Vanicek J: **Optimal use of conservation and accessibility filters in microRNA target prediction**. *PLoS One* 2012, 7(2):e32 208.

[205] Marín RM and Vaníček J: **Efficient use of accessibility in microRNA target prediction**. *Nucleic Acids Res* 2011, 39(1):19–29.

[206] Markham NR and Zuker M: **UNAFold: software for nucleic acid folding and hybridization**. *Methods Mol Biol* 2008, 453:3–31.

[207] Mathews D, Sabina J, Zuker M, and Turner D: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure**. *J Mol Biol* 1999, 288(5):911–40.

[208] Mathews DH and Turner DH: **Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops**. *Biochemistry* 2002, 41(3):869–80.

[209] Mattick JS, Taft RJ, and Faulkner GJ: **A global view of genomic information–moving beyond the gene and the master regulator**. *Trends in Genetics* 2010, 26(1):21–8.

[210] McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure**. *Biopolymers* 1990, 29(6-7):1105–19.

[211] Meister G: **Argonaute proteins: functional insights and emerging roles**. *Nat Rev Genet* 2013, 14(7):447–59.

[212] Mendell JT and Olson EN: **MicroRNAs in stress signaling and human disease**. *Cell* 2012, 148(6):1172–87.

[213] Mendes ND, Freitas AT, and Sagot MF: **Current tools for the identification of miRNA genes and their targets**. *Nucleic Acids Res* 2009, 37(8):2419–33.

[214] Milek M, Wyler E, and Landthaler M: **Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing**. *Semin Cell Dev Biol* 2012, 23(2):206–12.

[215] Min H and Yoon S: **Got target? Computational methods for microRNA target prediction and their extension**. *Exp Mol Med* 2010, 42(4):233–44.

[216] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, and Rigoutsos I: **A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes**. *Cell* 2006, 126(6):1203–17.

[217] Mitra R and Bandyopadhyay S: **MultiMiTar: a novel multi objective optimization based miRNA-target prediction method**. *PLoS One* 2011, 6(9):e24 583.

[218] Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J, *et al.*: **An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803**. *Proc Natl Acad Sci USA* 2011, 108(5):2124–9.

[219] Möhl M, Salari R, Will S, Backofen R, and Sahinalp SC: **Sparsification of RNA structure prediction including pseudoknots**. *Algorithms Mol Biol* 2010, 5(1):39.

[220] Moreno R, Marzi S, Romby P, and Rojo F: **The Crc global regulator binds to an unpaired A-rich motif at the Pseudomonas putida alkS mRNA coding sequence and inhibits translation initiation**. *Nucleic Acids Res* 2009, 37(22):7678–90.

[221] Morey M, Fernandez-Marmiesse A, Castineiras D, Fraga JM, Couce ML, and Cocho JA: **A glimpse into past, present, and future DNA sequencing**. *Mol Genet Metab* 2013, 110(1-2):3–24.

[222] Morgan SR and Higgs PG: **Evidence for kinetic effects in the folding of large RNA molecules**. *The Journal of Chemical Physics* 1996, 105(16):7152.

[223] Mortimer SA and Weeks KM: **Time-resolved RNA SHAPE chemistry: quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution**. *Nat Protoc* 2009, 4(10):1413–21.

[224] Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, and Hofacker IL: **Thermodynamics of RNA-RNA binding**. *Bioinformatics* 2006, 22(10):1177–82.

[225] Muniategui A, Pey J, Planes FJ, and Rubio A: **Joint analysis of miRNA and mRNA expression data**. *Brief Bioinform* 2013, 14(3):263–78.

[226] Nakanishi K, Weinberg DE, Bartel DP, and Patel DJ: **Structure of yeast Argonaute with guide RNA**. *Nature* 2012, 486(7403):368–74.

[227] Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, and Ke A: **Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system**. *Structure* 2012, 20(9):1574–84.

[228] Nawrocki EP, Kolbe DL, and Eddy SR: **Infernal 1.0: inference of RNA alignments**. *Bioinformatics* 2009, 25(10):1335–7.

[229] Needleman SB and Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, 48(3):443–53.

[230] Needleman SB and Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, 48(3):443–53.

[231] Notredame C, Higgins DG, and Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, 302(1):205–17.

[232] Novikova IV, Hennelly SP, and Sanbonmatsu KY: **Tackling Structures of Long Noncoding RNAs**. *Int J Mol Sci* 2013, 14(12):23 672–84.

[233] Nussinov R and Tinoco IJ: **Sequential folding of a messenger RNA molecule**. *J Mol Biol* 1981, 151(3):519–33.

[234] Obernosterer G, Tafer H, and Martinez J: **Target site effects in the RNA interference and microRNA pathways**. *Biochem Soc Trans* 2008, 36(Pt 6):1216–9.

[235] Orom UA, Nielsen FC, and Lund AH: **MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation**. *Mol Cell* 2008, 30(4):460–71.

[236] Ossowski S, Schwab R, and Weigel D: **Gene silencing in plants using artificial microRNAs and other small RNAs**. *Plant J* 2008, 53(4):674–90.

[237] Ouyang Z, Snyder MP, and Chang HY: **SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data**. *Genome Res* 2013, 23(2):377–87.

[238] Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, and Blencowe BJ: **Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression**. *Genes Dev* 2006, 20(2):153–8.

[239] Papantonis A and Cook PR: **Transcription factories: genome organization and gene regulation**. *Chem Rev* 2013, 113(11):8683–705.

[240] Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, and Haussler D: **Identification and Classification of Conserved RNA Secondary Structures in the Human Genome**. *PLoS Comput Biol* 2006, 2(4):e33.

[241] Peng SS, Chen CY, Xu N, and Shyu AB: **RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein**. *EMBO J* 1998, 17(12):3461–70.

[242] Perez I, Lin CH, McAfee JG, and Patton JG: **Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo**. *RNA* 1997, 3(7):764–78.

[243] Peritz T, Zeng F, Kannanayakal TJ, Kilk K, Eiriksdottir E, Langel U, and Eberwine J: **Immunoprecipitation of mRNA-protein complexes**. *Nat Protoc* 2006, 1(2):577–80.

[244] Petrov AI, Zirbel CL, and Leontis NB: **Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas**. *RNA* 2013, 19(10):1327–40.

[245] Pratt AJ and MacRae IJ: **The RNA-induced silencing complex: a versatile gene-silencing machine**. *Journal of Biological Chemistry* 2009, 284(27):17 897–901.

[246] Ragan C, Cloonan N, Grimmond SM, Zuker M, and Ragan MA: **Transcriptome-wide prediction of miRNA targets in human and mouse using FASTH**. *PLoS ONE* 2009, 4(5):e5745.

[247] Ragoussis J, Elvidge GP, Kaur K, and Colella S: **Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research**. *PLoS Genet* 2006, 2(7):e100.

[248] Randau L: **RNA processing in the minimal organism Nanoarchaeum equitans**. *Genome Biol* 2012, 13(7):R63.

[249] Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, *et al.*: **A compendium of RNA-binding motifs for decoding gene regulation**. *Nature* 2013, 499(7457):172–7.

[250] Reczko M, Maragkakis M, Alexiou P, Grosse I, and Hatzigeorgiou AG: **Functional microRNA targets in protein coding sequences**. *Bioinformatics* 2012, 28(6):771–6.

[251] Reczko M, Maragkakis M, Alexiou P, Papadopoulos GL, and Hatzigeorgiou AG: **Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data**. *Front Genet* 2011, 2:103.

[252] Rehmsmeier M, Steffen P, Höchsmann M, and Giegerich R: **Fast and effective prediction of microRNA/target duplexes**. *RNA* 2004, 10(10):1507–17.

[253] Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF, and Nebel ME: **Topology and prediction of RNA pseudoknots**. *Bioinformatics* 2011, 27(8):1076–85.

[254] Reinharz V, Major F, and Waldispuhl J: **Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure**. *Bioinformatics* 2012, 28(12):i207–i214.

[255] Reuter JS and Mathews DH: **RNAstructure: software for RNA secondary structure prediction and analysis**. *BMC Bioinformatics* 2010, 11:129.

[256] Rho M, Wu YW, Tang H, Doak TG, and Ye Y: **Diverse CRISPRs evolving in human microbiomes**. *PLoS Genet* 2012, 8(6):e1002 441.

[257] Richter AS: *Computational analysis and prediction of RNA-RNA interactions*. Ph.D. thesis, University of Freiburg, 2012.

[258] Richter AS and Backofen R: **Accessibility and conservation: General features of bacterial small RNA-mRNA interactions?** *RNA Biol* 2012, 9(7):954–65.

[259] Richter AS, Schleberger C, Backofen R, and Steglich C: **Seed-based IntaRNA prediction combined with GFP-reporter system identifies mRNA targets of the small RNA Yfr1**. *Bioinformatics* 2010, 26(1):1–5.

[260] Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, and Randau L: **Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis**. *Nucleic Acids Res* 2012, 40(19):9887–96.

[261] Rife JP, Stallings SC, Correll CC, Dallas A, Steitz TA, and Moore PB: **Comparison of the crystal and solution structures of two RNA oligonucleotides**. *Biophys J* 1999, 76(1 Pt 1):65–75.

[262] Ritchie W, Rajasekhar M, Flamant S, and Rasko JEJ: **Conserved expression patterns predict microRNA targets**. *PLoS Comput Biol* 2009, 5(9):e1000 513.

[263] Rivas E and Eddy SR: **The language of RNA: a formal grammar that includes pseudoknots**. *Bioinformatics* 2000, 16(4):334–40.

[264] Roberts L and Holcik M: **RNA structure: new messages in translation, replication and disease. Workshop on the role of RNA structures in the translation of viral and cellular RNAs**. *EMBO Rep* 2009, 10(5):449–53.

[265] Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, and Mesirov JP: **Integrative genomics viewer**. *Nat Biotechnol* 2011, 29(1):24–6.

[266] Rose D, Hackermuller J, Washietl S, Reiche K, Hertel J, Findeiss S, Stadler PF, and Prohaska SJ: **Computational RNomics of drosophilids**. *BMC Genomics* 2007, 8:406.

[267] Rose D, Joris J, Hackermuller J, Reiche K, Li Q, and Stadler PF: **Duplicated RNA genes in teleost fish genomes**. *J Bioinform Comput Biol* 2008, 6(6):1157–75.

[268] Rouskin S, Zubradt M, Washietl S, Kellis M, and Weissman JS: **Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo**. *Nature* 2014, 505(7485):701–5.

[269] Rousseau C, Gonnet M, Le Romancer M, and Nicolas J: **CRISPI: a CRISPR interactive database**. *Bioinformatics* 2009, 25(24):3317–8.

[270] Ruhl C, Stauffer E, Kahles A, Wagner G, Drechsel G, Ratsch G, and Wachter A: **Polypyrimidine tract binding protein homologs from Arabidopsis are key regulators of alternative splicing with implications in fundamental developmental processes**. *Plant Cell* 2012, 24(11):4360–75.

[271] Saetrom P, Heale BSE, Snove OJ, Aagaard L, Alluin J, and Rossi JJ: **Distance constraints between microRNA target sites dictate efficacy and cooperativity**. *Nucleic Acids Res* 2007, 35(7):2333–42.

[272] Salmena L, Poliseno L, Tay Y, Kats L, and Pandolfi PP: **A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language?** *Cell* 2011.

[273] Sankoff D: **Simultaneous solution of the RNA folding, alignment and protosequence problems.** *SIAM J Appl Math* 1985, 45(5):810–825.

[274] Sashital DG, Jinek M, and Doudna JA: **An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3**. *Nat Struct Mol Biol* 2011, 18(6):680–7.

[275] Sawicka K, Bushell M, Spriggs KA, and Willis AE: **Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein**. *Biochem Soc Trans* 2008, 36(Pt 4):641–7.

[276] Schmitter D, Filkowski J, Sewer A, Pillai RS, Oakeley EJ, Zavolan M, Svoboda P, and Filipowicz W: **Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells**. *Nucleic Acids Res* 2006, 34(17):4801–15.

[277] Schnall-Levin M, Zhao Y, Perrimon N, and Berger B: **Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3'UTRs**. *Proc Natl Acad Sci USA* 2010, 107(36):15 751–6.

[278] Schueler M, Munschauer M, Gregersen LH, Finzel A, Loewer A, Chen W, Landthaler M, and Dieterich C: **Differential protein occupancy profiling of the mRNA transcriptome**. *Genome Biol* 2014, 15(1):R15.

[279] Schwanhausser B, Gossen M, Dittmar G, and Selbach M: **Global analysis of cellular protein translation by pulsed SILAC**. *Proteomics* 2009, 9(1):205–9.

[280] Scott LG and Hennig M: **RNA structure determination by NMR**. *Methods Mol Biol* 2008, 452:29–61.

[281] Seemann SE, Gorodkin J, and Backofen R: **Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments**. *Nucleic Acids Res* 2008, 36(20):6355–62.

[282] Seemann SE, Richter AS, Gesell T, Backofen R, and Gorodkin J: **PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences**. *Bioinformatics* 2011, 27(2):211–219.

[283] Seemann SE, Richter AS, Gorodkin J, and Backofen R: **Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions**. *Algorithms Mol Biol* 2010, 5:22.

[284] Seetin MG and Mathews DH: **TurboKnot: rapid prediction of conserved RNA secondary structures including pseudoknots**. *Bioinformatics* 2012, 28(6):792–8.

[285] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, and Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs**. *Nature* 2008, 455(7209):58–63.

[286] Shah SA, Erdmann S, Mojica FJM, and Garrett RA: **Protospacer recognition motifs: Mixed identities and functional diversity**. *RNA Biol* 2013, 10(5).

[287] Shah SA and Garrett RA: **CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems**. *Res Microbiol* 2011, 162(1):27–38.

[288] Shao Y, Wu Y, Chan CY, McDonough K, and Ding Y: **Rational design and rapid screening of antisense oligonucleotides for prokaryotic gene modulation**. *Nucleic Acids Res* 2006, 34(19):5660–9.

[289] Silverman SK, Zheng M, Wu M, Tinoco IJ, and Cech TR: **Quantifying the energetic interplay of RNA tertiary and secondary structure interactions**. *RNA* 1999, 5(12):1665–74.

[290] Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, and Siksnys V: **In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus**. *EMBO J* 2013, 32(3):385–94.

[291] Smith C, Heyne S, Richter AS, Will S, and Backofen R: **Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA**. *Nucleic Acids Res* 2010, 38 Suppl:W373–7.

[292] Smith T and Waterman M: **Comparison of biosequences**. *Adv appl Math* 1981, 2:482–489.

[293] Soifer HS, Rossi JJ, and Saetrom P: **MicroRNAs in disease and potential therapeutic applications**. *Mol Ther* 2007, 15(12):2070–9.

[294] Soon WW, Hariharan M, and Snyder MP: **High-throughput sequencing for biology and medicine**. *Mol Syst Biol* 2013, 9:640.

[295] Staals RHJ, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K, *et al.*: **Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus**. *Mol Cell* 2013, 52(1):135–45.

[296] Steffen P, Voss B, Rehmsmeier M, Reeder J, and Giegerich R: **RNAshapes: an integrated RNA analysis package based on abstract shapes**. *Bioinformatics* 2006, 22(4):500–3.

[297] Stefl R, Skrisovska L, and Allain FHT: **RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle**. *EMBO Rep* 2005, 6(1):33–8.

[298] Sternberg SH, Haurwitz RE, and Doudna JA: **Mechanism of substrate selection by a highly specific CRISPR endoribonuclease**. *RNA* 2012, 18(4):661–72.

[299] Sturm M, Hackenberg M, Langenberger D, and Frishman D: **TargetSpy: a supervised machine learning approach for microRNA target prediction**. *BMC Bioinformatics* 2010, 11:292.

[300] Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, and Ule J: **Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions**. *Genome Biol* 2012, 13(8):R67.

[301] Sztuba-Solinska J and Le Grice SFJ: **Probing retroviral and retrotransposon genome structures: The "SHAPE" of things to come**. *Mol Biol Int* 2012, 2012:530 754.

[302] Tacke R, Chen Y, and Manley JL: **Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer**. *Proc Natl Acad Sci USA* 1997, 94(4):1148–53.

[303] Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, and Hofacker IL: **The impact of target site accessibility on the design of effective siRNAs**. *Nat Biotechnol* 2008, 26(5):578–83.

[304] Tafer H and Hofacker IL: **RNAplex: a fast tool for RNA-RNA interaction search**. *Bioinformatics* 2008, 24(22):2657–63.

[305] Tarang S and Weston MD: **Macros in microRNA target identification: A comparative analysis of in silico, in vitro, and in vivo approaches to microRNA target identification**. *RNA Biol* 2014, 11(4):324–333.

[306] Tay Y, Zhang J, Thomson AM, Lim B, and Rigoutsos I: **MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation**. *Nature* 2008, 455(7216):1124–8.

[307] Terns MP and Terns RM: **CRISPR-based adaptive immune systems**. *Curr Opin Microbiol* 2011, 14(3):321–7.

[308] Terns RM and Terns MP: **CRISPR-based technologies: prokaryotic defense weapons repurposed**. *Trends in Genetics* 2014, 30(3):111–118.

[309] Theocharidis A, van Dongen S, Enright AJ, and Freeman TC: **Network visualization and analysis of gene expression data using BioLayout Express(3D)**. *Nat Protoc* 2009, 4(10):1535–50.

[310] Thomas M, Lieberman J, and Lal A: **Desperately seeking microRNA targets**. *Nat Struct Mol Biol* 2010, 17(10):1169–74.

[311] Thompson JD, Higgins DG, and Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, 22(22):4673–80.

[312] Tinoco IJ and Bustamante C: **How RNA folds**. *J Mol Biol* 1999, 293(2):271–81.

[313] Tippmann SC, Ivanek R, Gaidatzis D, Scholer A, Hoerner L, van Nimwegen E, Stadler PF, Stadler MB, and Schubeler D: **Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels**. *Mol Syst Biol* 2012, 8:593.

[314] Tjaden B: **TargetRNA: a tool for predicting targets of small RNA action in bacteria**. *Nucleic Acids Res* 2008, 36(Web Server issue):W109–13.

[315] Turner DH and Mathews DH: **NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure**. *Nucleic Acids Res* 2010, 38(Database issue):D280–2.

[316] Ule J, Jensen K, Mele A, and Darnell RB: **CLIP: a method for identifying protein-RNA interaction sites in living cells**. *Methods* 2005, 37(4):376–86.

[317] Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, and Haussler D: **FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing**. *Nat Methods* 2010, 7(12):995–1001.

[318] Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, Hannon GJ, Sanford JR, Penalva LOF, and Smith AD: **Site identification in high-throughput RNA-protein interaction data**. *Bioinformatics* 2012, 28(23):3013–20.

[319] van Bakel H, Nislow C, Blencowe BJ, and Hughes TR: **Most "dark matter" transcripts are associated with known genes**. *PLoS Biol* 2010, 8(5):e1000 371.

[320] van Dongen S: *Graph Clustering by Flow Simulation.* Ph.D. thesis, University of Utrecht, The Netherlands, 2000.

[321] Vasudevan S, Tong Y, and Steitz JA: **Switching from repression to activation: microRNAs can up-regulate translation**. *Science* 2007, 318(5858):1931–4.

[322] Veerla S and Hoglund M: **Analysis of promoter regions of co-expressed genes identified by microarray analysis**. *BMC Bioinformatics* 2006, 7:384.

[323] Vejnar CE and Zdobnov EM: **MiRmap: comprehensive prediction of microRNA target repression strength**. *Nucleic Acids Res* 2012, 40(22):11 673–83.

[324] Vestergaard G, Garrett RA, and Shah SA: **CRISPR adaptive immune systems of Archaea**. *RNA Biol* 2014, 11(2):157–168.

[325] Vogel J and Wagner EGH: **Target identification of small noncoding RNAs in bacteria**. *Curr Opin Microbiol* 2007, 10(3):262–70.

[326] Wang R, Preamplume G, Terns MP, Terns RM, and Li H: **Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage**. *Structure* 2011, 19(2):257–64.

[327] Wang R, Zheng H, Preamplume G, Shao Y, and Li H: **The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex**. *Protein Sci* 2012, 21(3):405–17.

[328] Wang T, Wei JJ, Sabatini DM, and Lander ES: **Genetic screens in human cells using the CRISPR-Cas9 system**. *Science* 2014, 343(6166):80–4.

[329] Wang X, Juan L, Lv J, Wang K, Sanford JR, and Liu Y: **Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1**. *BMC Genomics* 2011, 12 Suppl 5:S8.

[330] Wang Y, Juranek S, Li H, Sheng G, Tuschl T, and Patel DJ: **Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex**. *Nature* 2008, 456(7224):921–6.

[331] Warthmann N, Ossowski S, Schwab R, and Weigel D: **Artificial microRNAs for specific gene silencing in rice**. *Methods Mol Biol* 2013, 956:131–49.

[332] Washietl S and Hofacker IL: **Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics**. *J Mol Biol* 2004, 342(1):19–30.

[333] Washietl S, Hofacker IL, and Stadler PF: **Fast and reliable prediction of noncoding RNAs**. *Proc Natl Acad Sci USA* 2005, 102(7):2454–9.

[334] Washietl S, Hofacker IL, Stadler PF, and Kellis M: **RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction**. *Nucleic Acids Res* 2012, 40(10):4261–72.

[335] Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, *et al.*: **Structured RNAs in the ENCODE selected regions of the human genome**. *Genome Res* 2007, 17(6):852–64.

[336] Weeks KM: **RNA structure probing dash seq**. *Proc Natl Acad Sci USA* 2011, 108(27):10 933–4.

[337] Wiedenheft B, Sternberg SH, and Doudna JA: **RNA-guided genetic silencing systems in bacteria and archaea**. *Nature* 2012, 482(7385):331–8.

[338] Will S, Joshi T, Hofacker IL, Stadler PF, and Backofen R: **LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs**. *RNA* 2012, 18(5):900–14.

[339] Will S, Reiche K, Hofacker IL, Stadler PF, and Backofen R: **Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering**. *PLoS Comput Biol* 2007, 3(4):e65.

[340] Will S, Siebauer MF, Heyne S, Engelhardt J, Stadler PF, Reiche K, and Backofen R: **LocARNAscan: Incorporating thermodynamic stability in sequence and structure-based RNA homology search**. *Algorithms Mol Biol* 2013, 8(1):14.

[341] Will S, Yu M, and Berger B: **Structure-based whole-genome realignment reveals many novel noncoding RNAs**. *Genome Res* 2013, 23(6):1018–1027.

[342] Witkos TM, Koscianska E, and Krzyzosiak WJ: **Practical Aspects of microRNA Target Prediction**. *Curr Mol Med* 2011, 11(2):93–109.

[343] Wright PR, Richter AS, Papenfort K, Mann M, Vogel J, Hess WR, Backofen R, and Georg J: **Comparative genomics boosts target prediction for bacterial small RNAs**. *Proc Natl Acad Sci USA* 2013, 110(37):E3487–96.

[344] Wuchty S, Fontana W, Hofacker IL, and Schuster P: **Complete suboptimal folding of RNA and the stability of secondary structures**. *Biopolymers* 1999, 49(2):145–65.

[345] Xia Z, Clark P, Huynh T, Loher P, Zhao Y, Chen HW, Rigoutsos I, and Zhou R: **Molecular dynamics simulations of Ago silencing complexes reveal a large repertoire of admissible 'seed-less' targets**. *Sci Rep* 2012, 2:569.

[346] Xiao B, Li W, Guo G, Li B, Liu Z, Jia K, Guo Y, Mao X, and Zou Q: **Identification of small noncoding RNAs in Helicobacter pylori by a bioinformatics-based approach**. *Curr Microbiol* 2009, 58(3):258–63.

[347] Yang JH, Li JH, Shao P, Zhou H, Chen YQ, and Qu LH: **starbase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data**. *Nucleic Acids Res* 2011, 39(Database issue):D202–9.

[348] Yang JH and Qu LH: **Discovery of microRNA regulatory networks by integrating multidimensional high-throughput data**. *Adv Exp Med Biol* 2013, 774:251–66.

[349] Yao Z, Weinberg Z, and Ruzzo WL: **CMfinder – a covariance model based RNA motif finding algorithm**. *Bioinformatics* 2006, 22(4):445–52.

[350] Yousef M, Jung S, Showe LC, and Showe MK: **Learning from positive examples when the negative class is undetermined–microRNA gene identification**. *Algorithms Mol Biol* 2008, 3:2.

[351] Yusupova GZ, Yusupov MM, Cate JH, and Noller HF: **The path of messenger RNA through the ribosome**. *Cell* 2001, 106(2):233–41.

[352] Zhang C and Darnell RB: **Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data**. *Nat Biotechnol* 2011, 29(7):607–14.

[353] Zhang C and Galbraith DW: **RNA interference-mediated gene knockdown within specific cell types**. *Plant Mol Biol* 2012, 80(2):169–76.

[354] Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV, *et al.*: **Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity**. *Mol Cell* 2012, 45(3):303–13.

[355] Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, and Yeo GW: **Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans**. *Nat Struct Mol Biol* 2010, 17(2):173–9.

[356] Zuker M and Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information**. *Nucleic Acids Res* 1981, 9(1):133–48.