TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Study of seed-based microRNA targeting in mammals

Florian A. Büttner

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. B. Küster
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. H.-W. Mewes
	2. Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 07.08.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 22.11.2013 angenommen.

To Katrin and Paulo

Danksagung

Zuerst möchte ich mich bei meinem Doktorvater Prof. Dr. Hans-Werner Mewes bedanken, mir diese Doktorarbeit ermöglicht zu haben, sowie für die Betreuung und wohlwollende Unterstützung in den letzten Jahren.

Für stimulierende und spannende Zusammenarbeit und Diskussionen möchte ich mich vor allem bei Daniel Ellwanger und Florian Giesert, sowie bei Sebastian Toepel, Matthias Arnold, Dominik Lutter und Jörn Leonhardt bedanken.

Weiter geht mein Dank an alle Kollegen vom Institut für Bioinformatik und Systembiologie und vom Lehrstuhl für Genomorientierte Bioinformatik in Weihenstephan für das freundliche, kollegiale und unterhaltsame Miteinander.

Erwähnung finden müssen an dieser Stelle auch meine Freunde und Bekannten außerhalb von IBIS und Universität. Die wöchentlichen Fußballabende waren einfach Weltklasse!

Herzlich danken möchte ich meinen Eltern, meiner Schwester Barbara und meinem Bruder Tobias für die vielen Formen der Unterstützung, die sie mir haben zukommen lassen.

Mein besonderer Dank geht an Katrin und Paulo, eure Unterstützung und euer Verständnis waren von unschätzbarem Wert für mich in den letzten Jahren!

Abstract

microRNAs or miRNAs represent an important class of small non-coding RNAs that post-transcriptionally regulate gene expression in almost all biological processes. miRNAs serve as specificity factors that guide the gene silencing ribonucleoprotein complex to the transcript to be regulated. The basic requirement for miRNA target recognition is a perfect match of 6-8 nucleotides (nt) between the seed region located at the 5' end of the miRNA and the target.

Recent biochemical approaches for the transcriptome-wide identification of interactions between miRNAs and targets revealed that miRNAs likewise pair with sites located in the 3'-untranslated region (3'UTR) and the coding region disproving the previous notion stating that miRNAs act mainly through the 3'UTR. Using experimentally identified miRNA target sites, miRNA seed-based targeting has been thoroughly analyzed in this thesis. We found that in both mRNA regions most miRNA-target interactions are based on short 6 nt long and less conserved seed matches. As short seed matches affect target expression less compared to long matches, the hypothesis was proposed that the main function of short sites is to operate as miRNA regulatory elements that prevent efficient regulation of other targets containing long seed matches by sponging miRNAs.

Second, an evaluation of representative and frequently used miRNA target prediction programs was conducted. In contrast to previous comparisons of methods that considered the prediction of miRNA-target interactions, the focus of this analysis was to assess the ability of predictions methods to identify the location of the site of miRNA-target interaction within the target sequence. Although most of the methods initially search for short and long seed matches, among the top ranked predictions long seed matches were greatly over-represented. This analysis revealed that prediction methods are still far away from a comprehensive determination of miRNA target sites. Combining the output of different algorithms turned out to be a promising approach to increase sensitivity.

Zusammenfassung

microRNAs oder miRNAs stellen eine bedeutende Klasse kurzer, nicht kodierender RNAs dar, die Genexpression in fast allen biologischen Prozessen posttranskriptional regulieren. miRNAs gewährleisten die Spezifität bei der Regulation, indem sie den Ribonucleoproteinkomplex, der die Repression der Genexpression vollzieht, zur Ziel-mRNA führen. Ausschlaggebend für die Ziel-mRNA Erkennung durch die miRNA ist die perfekte Komplementarität zwischen einer 6-8 Nukleotide langen Sequenz, die sich in der Seed-Region am 5' Ende der miRNA befindet, und der Seed-Bindesequenz auf der Targetsequenz.

Moderne biochemische Methoden zur Transkriptom-weiten Bestimmung von miRNA-Target Interaktionen deckten auf, dass miRNAs gleichermaßen an Bindestellen im 3'UTR sowie in der kodierenden Region binden und widerlegten damit die bis dahin gängige Auffassung, dass miRNAs hauptsächlich mit dem 3'UTR interagieren. Unter Verwendung experimentell bestimmter miRNA-Bindestellen wurde in dieser Arbeit die auf der Seed-Region basierende Targeterkennung analysiert. Es zeigte sich, dass in beiden mRNA-Regionen die meisten miRNA-Ziel-mRNA Interaktionen auf kurzen 6 Nukleotide langen and wenig konservierten Seed-Bindesequenzen beruhen. Da kurze Seed-Bindesequenzen im Vergleich zu langen die Expression der Ziel-mRNA nur gering beeinflussen, wurde die Hypothese aufgestellt, dass sie in ihrer Hauptfunktion als miRNA regulierende Elemente fungieren. Die Interaktion mit kurzen Bindesequenzen hält miRNAs davon ab andere Targets mit langen Seed-Bindesequenzen zu regulieren.

Desweiteren wurde eine Gruppe repräsentativer und häufig genutzter Target-Vorhersagealgorithmen für miRNAs evaluiert. Im Gegensatz zu vorhergehenden Methodenvergleichen, die nur die miRNA-Target Interaktion berücksichtigten, wurde hier die Fähigkeit beurteilt, die Position der Seed-Bindesequenz innerhalb der Target-Sequenz zu detektieren. Zwar suchen die meisten Methoden kurze und lange Bindesequenzen, jedoch sind lange Bindesequenzen unter den höchst bewerteten Vorhersagen deutlich überrepräsentiert. Gegenwärtig sind Target-Vorhersagemethoden noch weit davon entfernt, umfassend und präzise miRNA Bindestellen aufzuspüren. Die Kombination der Resultate verschiedener Methoden stellte sich jedoch als vielversprechender Ansatz zur Erhöhung der Sensitivität heraus.

List of Abbreviations

3'UTR	3'-untranslated region
4SU	photoreactive 4-thiouridine
5'UTR	5'-untranslated region
Ago	Argonaute
C. elegans	*Caenorhabditis elegans*
CCR	crosslink-centered region, page 36
CCRx	CCR with T/C mutation on codon position x
cDNA	complementary DNA
ceRNA	competing endogenous RNA, page 33
CLIP	crosslinking and IP
eIF4F	eukaryotic translation-initiation factor
HITS-CLIP	High-throughput sequencing of RNA isolated by CLIP, page 35
IBIS	Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München
IDG	Institute of Developmental Genetics at the Helmholtz Zentrum München
IP	immunoprecipitation
LOR	log odds ratio
Lrrk2	Leucine-rich repeat kinase 2
MCC	Matthews correlation coefficient

nt	nucleotide(s)
OR	odds ratio
ORF	open reading frame
PABP	poly(A)-binding protein
PAR-CLIP	photoactivatable-ribonucleoside-enhanced CLIP, page 36
PCC	Pearson correlation coefficient
pri-miRNA	primary miRNA transcript
PSC	precision-sensitivity curve
R^2	squared PCC
RBP	RNA-binding protein
RefSeq	NCBI reference sequence
RISC	RNA-induced-silencing complex
RNAi	RNA interference
stRNA	small temporal RNA
TRBP	transactivation-responsive RBP
UV	ultraviolet
WC	Watson-Crick

List of Figures

List of Tables

Contents

1 Introduction

1.1 Motivation

The first miRNA, lin-4, was found in 1993 by a collaboration of the laboratories of Victor Ambros and Gary Ruvkun. lin-4 downregulates post-transcriptionally the expression of the protein-coding gene lin-14 during development of the nematode *Caenorhabditis elegans* (*C. elegans*) [2],[3]. The studies of the groups of Tom Tuschl [4], David Bartel [5] and Victor Ambros [6], presented in the same issue of *Science* eight years later, demonstrated that post-transcriptional gene regulation through miRNAs is not a special feature only appearing during development of nematodes but instead a conserved and prevalent regulatory mechanism. Naturally, this discovery raised a number of questions, e.g.: How many miRNAs are there? How are miRNAs generated? How is miRNA biogenesis regulated? How do miRNAs regulate their targets? How many miRNA targets are there? How do miRNAs identify their targets? What role do miRNAs play in gene-regulatory pathways?

The last two questions were of particular interest in our group. Our aim was to model qualitative systems biological networks with particular focus on the post-transcriptional regulation by miRNAs. One way to integrate miRNAs in biological models would have been to use existing miRNA target prediction algorithms. But soon it became apparent that these methods generate very large numbers of predictions. Further, the intersection of the sets of predictions of different methods was relatively low [7]. The search for genuine miRNA targets is exacerbated by the shortness of the miRNA recognition elements on the target sequences. Only a 6-8 nt long sub-sequence located at the 5' end of the miRNA pairs perfectly to the target sequence in animals. The remainder of the $\sim$ 22 nt long miRNA base-pairs less regularly. The myriads of target predictions and the small overlap between the prediction tools motivated us to carry out an analysis of miRNA targeting to be able to assess better the reliability of both target predictions and target prediction methods.

For our analysis we used miRNA target sites that were detected by HITS-CLIP [8] and PAR-CLIP [9]. These are biochemical approaches for the transcriptome-wide identification of miRNA-target interactions. Using experimentally verified

miRNA target sites we have classified and characterized different types of miRNA seed matching patterns, subsequently called seed types. In contrast to previous classifications of this kind, see [7], that were restricted to conserved target sites, we considered both conserved and species-specific target sites.

One important finding of both CLIP studies was that coding regions include a substantial proportion of target sites. Previously, the interest was focused almost exclusively on the 3'UTR. There, the first target sites were found and the repressive effect through sites located in the 3'UTR is generally stronger. In addition, initially it appeared unlikely that the coding region encodes beside the highly conserved genetic code further signals in large scale. In this thesis, miRNA targeting in both regions was investigated and compared with each other. In the second part, representative and frequently used target prediction methods were assessed. Particularly, the consideration of the various seed types by the tools was quantified.

1.2 Overview of this thesis

The following briefly summarizes the content of the individual chapters of this thesis.

Chapter 2 introduces the field of post-transcriptional gene regulation by miRNAs. After pointing out the importance of gene expression in eukaryotes in general, it will be focused on the impact of miRNA-mediated regulation on the protein output. Further, metabolism and regulation of the miRNA pathway are delineated. This chapter concludes with a description of the current molecular biological model of regulation through miRNAs.

In **Chapter 3**, the results of the analysis of miRNA seed-based targeting in 3'UTR and coding region are presented. This study builds up on a project with Daniel Ellwanger that has been published as *The sufficient minimal set of miRNA seed types* [1]. At the beginning, the current knowledge of target recognition by miRNAs is outlined. miRNA seed types were defined based on the miRNA-target interaction map provided by Hafner et al. [9]. The obtained seed types were characterized by means of quantitative and qualitative features such as specificity, conservation of seed matches, effect on mRNA stability and different preferences of miRNAs for seed types. Finally, a hypothesis was formulated suggesting that the main function of short seed matches is to operate as miRNA regulatory elements that prevent regulation of other targets by sponging miRNAs.

Chapter 4 includes the results of a performance evaluation of six miRNA target prediction methods. This chapter starts with a review of computational miRNA target prediction. Mainly, three aspects were considered within the evaluation.

First, it was questioned which seed types based on the classification in chapter 3 are searched by the tools. Further, it was examined how the various scoring functions differentiate between different seed types. Third, the performance of the prediction methods to correctly distinguish between true and false target sites was assessed.

In **Chapter 5** two ongoing projects are shortly introduced. In the analysis of miRNA targeting in the coding region it appeared that the occurrence of target sites is dependent of the reading frame. This observation and results of initial investigations are presented in the first section. The subsequent section introduces a project I am working on in collaboration with Florian Giesert from the Institute of Developmental Genetics (IDG) at the Helmholtz Zentrum München. We are exploring the post-transcriptional regulation of the Parkinson's disease-related Lrrk2 gene.

In **Chapter 6** the results presented in thesis are summarized and perspectives on possible future studies are given.

2 Background

2.1 Historical side note: The discovery of miRNAs

The importance of non-coding RNA for gene regulation emerged in 1993. *lin-4* and *lin-14* were the participants of the first regulatory relationship between a non-coding RNA and a mRNA to be discovered by a collaboration of the laboratories of Victor Ambros and Gary Ruvkun [2],[3].

lin-14 and *lin-4* are involved in developmental timing in *C. elegans*. Initially, Ambros and Ruvkun were working on the molecular identification of the gene *lin-14* [10]. *lin-14* controls the timing and sequence of many post-embryonic developmental events but is absent at later stages. Both *gain-of-function* and *loss-of-function* mutations were known for *lin-14* [11]. Early cell lineages with gain-of-function mutations are normal but later cells reiterate early development programs inappropriate for the larval stage resulting in a retarded phenotype. Loss-of-function mutations induced the opposite effect, i.e. a precocious development, suggesting that *lin-14* activity in the wild type is high in post-embryonic states and decreases for the proper expression of later cell fates [11]. It was hypothesized that *lin-4* is necessary for the decrease of *lin-14* activity, as the LIN-14 protein level remained high in later stages in animals with loss-of-function mutations in *lin-4*. Furthermore, the latter mutants exhibited a similar retarded phenotype as worms with gain-of-function mutations in *lin-14* [12].

The decoding of the *lin-14* gene structure revealed that the gain-of-function mutations were located in the 3'UTR. Fragments of the wild type 3'UTR were missing in the mutants [10],[13]. As the mutations did not affect the transcript level but only the protein expression of *lin-14*, a negative post-transcriptional regulation of *lin-14* by *lin-4* via cis-regulatory elements in the 3'UTR was suspected [13]. When Ambros and colleagues tried to clone the *lin-4* locus expecting a conventional protein, they were surprised not to find a coding sequence but a small 22 nt long non-coding RNA [2]. At the same time Ruvkun and colleagues found that *lin-14* is post-transcriptionally downregulated and that the 3'UTR of *lin-14* is necessary and sufficient for this regulation [3]. In particular, they detected seven conserved sites in the 3'UTR of high sequence similarity that were directly involved in regulation. The deletions in the mutant 3'UTR disrupted the effect of *lin-4*.

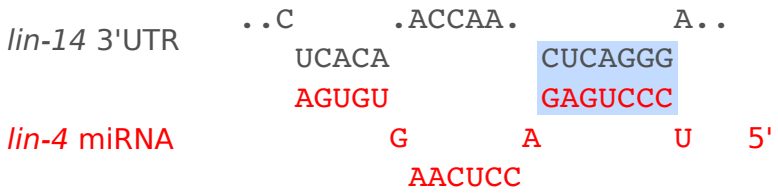


Figure 2.1: *lin-4-lin-14* DUPLEX. Predicted base-pairing of miRNA *lin-4* to target sites in the 3'UTR of *lin-14*. The miRNA seed match (see section 3.1.1) is highlighted in blue. This figure is based on a figure from [15].

Ambros and Ruvkun exchanged their unpublished sequence data and came to the same conclusion: The repeated sequence in the *lin-14* 3'UTR was partially complementary to the *lin-4* RNA [14], see figure 2.1. The first case of gene regulation mediated via a non-coding RNA had been identified.

Wickens and Takayama questioned in 1994 if the *lin-4-lin-14* type of gene regulatory interaction is a general mechanism [16] or if it is a "deviant" only occurring in the development of simple worms. Between 1993 and 2000 no further regulatory non-coding RNAs were discovered. Both Ambros and Ruvkun did not believe in a general role of small regulatory RNAs [14],[15], e.g. Ambros considered the *lin-4-lin-14* partnership a "nematode-specific curiosity" [15]. However, the finding of another target gene of *lin-4*, *lin-28*, in 1997 [17], the discovery of RNA interference (RNAi) [18] in 1998 or the identification of small antisense RNAs associated with RNAi in plants [19] in 1999 pointed to a possible generality of the mechanism exemplified by *lin-4*. But even the detection of the second small regulatory RNA, *let-7*, in *C. elegans* in 2000 by the Ruvkun laboratory [20] first did not dissuade Ambros and Ruvkun from assessing these RNAs to be singly components of nematode-specific developmental pathways. But when Pasquinelli et al. a coworker of Ruvkun showed that homologs of *let-7* exist across a broad range of animal phyla - including human - [21], the thinking changed instantly [15]. As both small RNAs, *lin-4* and *let-7*, are temporally expressed during *C. elegans* development, Pasquinelli et al. proposed to call this type of RNA *small temporal RNA* (stRNA). The new interest in small regulatory RNAs was further excited by the observation that regulatory relationships between stRNAs and targets are conserved [21],[22]. Eventually, like a gold rush the hunt after unknown stRNAs began. A first peak was reached in 2001 when the groups of Tuschl [4], Bartel [5] and Ambros [6] published their results in the same issue of *Science* each of them presenting numerous newly detected stRNAs. The terms *microRNA* or *miRNA* were coined in this issue as well [23].

2.2 Control of gene expression in eukaryotes

2.2.1 Importance of complex gene regulatory systems

Protein-coding sequences make up only a small fraction of a typical metazoan genome. In case of the human genome less than 1.5% is protein-coding. Before the human genome had been fully sequenced the estimates on the number of protein-coding genes exceeded the actual number of about 20,000 genes by several times. But unlike in prokaryotes where genome size and gene number are correlated, the vast majority of nuclear DNA is non-coding in eukaryotes [24]. Moreover, comparative analysis of sequenced genomes from various species indicated that organismal complexity is not reflected in the gene number in eukaryotes. For example, the genome of the simple worm *C. elegans* comprises about 19,000 genes [25], although the worm lacks the diversity of cell types and tissues seen in human, mouse ($\sim 25,000$ genes) [26] or even in the fruit fly *Drosophila melanogaster* (*D. melanogaster*) ($\sim 14,000$ genes) [27]. Therefore, understanding the reason for the broad spectrum of morphological and behavioral differences within the eukaryotic kingdom emerged as a major challenge in the early post-genomic era [28].

It became apparent that organismal complexity is correlated with the number of different gene expression patterns occurring within the life cycle of an eukaryote [29]. Gene expression is coordinated by various regulatory mechanisms that exert control at multiple stages on the pathway to the final protein. The multi-layered gene regulatory system involves cell signaling, chromatin modifications, transcriptional activation, post-transcriptional and post-translational regulations. Each single layer but also the cross-talk between regulatory layers is subject of investigation, see e.g. [30],[31],[32]. Nevertheless, most of the research efforts in this field so far have been invested in studying the regulation of transcription which represents the essential first step in gene expression [33]. This bias was due in part to historical reasons: Transcriptional activation had been the first regulatory mechanism to be discovered by Jacques Monod in prokaryotes [34]. On the other hand, biochemical methods applicable for its analysis such as microarray- and sequence-based technologies for large-scale mRNA quantification have been around longer. Further, the community was directed by the *central dogma of molecular biology* stating DNA is transcribed to RNA and RNA in turn is translated into protein. The central dogma largely holds true for prokaryotes, but in terms of animals the vast majority of DNA is neglected by this model [35]. Inter alia the discovery of regulatory non-coding RNA, especially miRNAs, leveraged an intensive engagement with post-transcriptional regulation in the last twenty years.

2.2.2 Post-transcriptional regulation

Proteins and their absolute concentrations, i.e. the final result of gene expression, are relevant to the phenotype. Protein production is controlled at multiple stages. Until recently, mainly transcriptional regulation has been in the focus [33]. But albeit necessary, transcriptional regulation is not sufficient to explain protein concentration completely, even not in less complex eukaryotes such as baker's yeast [36]. Particularly, the spatial and temporal uncoupling of transcription and translation in eukaryotes, i.e. transcription takes place in the nucleus and translation is conducted in the cytoplasm, gave rise to a wide domain for gene regulatory mechanisms performing post-transcriptional control - including mRNA processing, modification and decay as well as translation initiation, elongation, termination and protein degradation. Recent technological advances in mass spectrometry, RNA sequencing and microscopy enable simultaneous measurement of mRNA and protein concentrations [37]. Meanwhile, several studies have shown that significant discrepancies prevail between mRNA and protein levels, see e.g. [36],[38],[39].

Vogel et al. determined the average mRNA and protein levels for $> 1,000$ genes in a human tumor cell line [38]. They observed a significant but weak positive correlation between mRNA and protein concentration with a squared *Pearson correlation coefficient* (PCC), denoted R^2, of 0.29. R^2 reflects to what extent the variation in one variable can be explained by changes in another variable. Consequently, less than 30% of protein expression variance can be attributed to differences in mRNA expression suggesting substantial additional control at the level of translation. Vogel et al. analyzed the individual impact of ~ 200 sequence features on protein abundance. They considered features such as the lengths of the mRNA regions (coding and untranslated regions), nt and amino acid frequencies and properties and also potential miRNA target sites. Sequence lengths showed the strongest correlation, in particular coding sequence length is strongly inversely correlated but also long 3'UTRs weakened protein expression. Both observations were considered plausible: Precision of folding as well as of translation may decrease with the length of coding sequences. Long 3'UTRs provide more room for binding sites of potential negative regulators such as miRNAs or RNA-binding proteins (RBP) implying on the other hand short 3'UTRs to be accompanied by high protein abundance. It has been demonstrated in proliferating and in cancer cells that mRNAs were transcribed with shorter 3'UTRs leading to increased protein production due to a loss of cis-regulatory repressive elements such as miRNA target sites [40],[41]. Surprisingly, miRNA regulation itself, represented by predicted binding sites in 3'UTRs, did not significantly affect protein concentration in the analysis of Vogel et al.. Further, Vogel and colleagues determined the combined

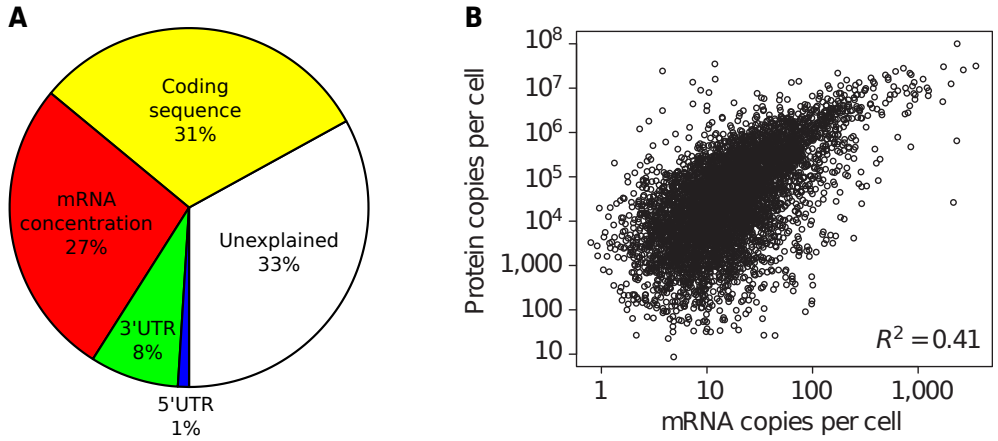


Figure 2.2: Relationship between mRNA and protein abundance. (A) Combining mRNA concentration and various sequence features of coding sequence and untranslated regions could explain 67% of variation of protein abundance for $> 1,000$ genes in a human tumor cell line. (B) Global quantification of absolute mRNA and protein abundance in mouse fibroblast cells showed that protein level variability could be attributed to 41% on variations of mRNA abundance. Figure A is taken from [38] and figure B is from [42].

contribution of mRNA expression, i.e. transcriptional regulation, mRNA stability, and sequence features on protein expression based on their data set, see figure 2.2(A). In this way 67% of protein abundance variation could be explained. Importantly, features related to the coding sequence constituted the greatest contributor with more than 30% revealing control at the level of translation and thereafter at least as important as regulation of mRNA transcription and mRNA stability. Referring to studies presented in section 2.2.3, Vogel et al. attributed the low impact of miRNA-mediated control to the presumable role of miRNAs to act primarily as fine-tuners of gene expression.

Schwanhäusser et al. globally quantified mRNA and protein expression levels and turnover in parallel in murine cells [39]. Their measurement of correlation between mRNA and protein concentration yielded a R^2 of 0.41, see figure 2.2(B). The features quantified by Schwanhäusser et al. allowed them to predict variations of the protein level almost completely. Most of variation could be attributed to changes in the translation rate ($R^2 \sim 0.5$). Concluding, the studies of Vogel et al. and Schwanhäusser et al. clearly point to a significant role of translational control during gene expression. According to the latter, regulation at the stage of translation is even dominating the protein level. However, one has to consider

that in both studies gene expression has been analyzed under steady-state conditions. Further, in both cases the expression values represented averages across cell populations. As reviewed by Vogel and Marcotte [37], methods to study the steps of gene expression on single cell level and in dynamic systems are emerging and will add further (unexpected) insights into the complex relationship between transcriptome and proteome.

2.2.3 Effect of miRNA-mediated control on protein abundance

miRNAs intervene in gene expression in two different ways either by repressing translation or by degradation of the target. Initially, the former mode had been considered to be the only one in animals due to the findings in *C. elegans*, see section 2.1. By contrast, plant miRNAs were thought to trigger cleavage and degradation of the targeted mRNA. This difference was attributed to the fact that in animals miRNAs base-pair only partially with their targets while in plants the entire miRNA binds to the target sequence. Meanwhile, it became clear that both modes of regulation occur in animals as well as in plants [43]. The mechanisms of miRNA regulation are described in section 2.4, here, the impact of miRNA-mediated regulation on protein abundance is outlined.

In two genome-wide studies the extent of control exerted by miRNAs was explored using quantitative mass spectrometry approaches to measure the protein level [44],[45]. Both groups examined the consequences triggered by both transfection and depletion of miRNAs and came to the conclusion that miRNAs affect protein production only mildly. Changes of expression levels were less than twofold suggesting miRNAs to fine-tune the protein level in general. Further, both reported that a single miRNA dampens the levels of hundreds of proteins confirming preceding computational predictions. It is important to note that these results were obtained from population averages which may obscure possibly stronger responses occurring in single cells. For example, the effect of lin-4 on lin-14 is rather switch-like than subtle.

Complementing the work of Selbach et al. and Baek et al., Mukherji and colleagues analyzed the effects of miRNA-mediated regulation in single cells [46]. Using quantitative fluorescence microscopy they measured the protein level in dependence of increasing mRNA concentration and constant miRNA level. They observed that protein production was not visible until a certain level of mRNA abundance was achieved. Above this threshold, the protein level increased linearly, see figure 2.3(A). Further, the threshold could be shifted by modulating the miRNA abundance and the sharpness of the threshold could be modified by the number of miRNA target sites, see figure 2.3(B). Based on these observations they

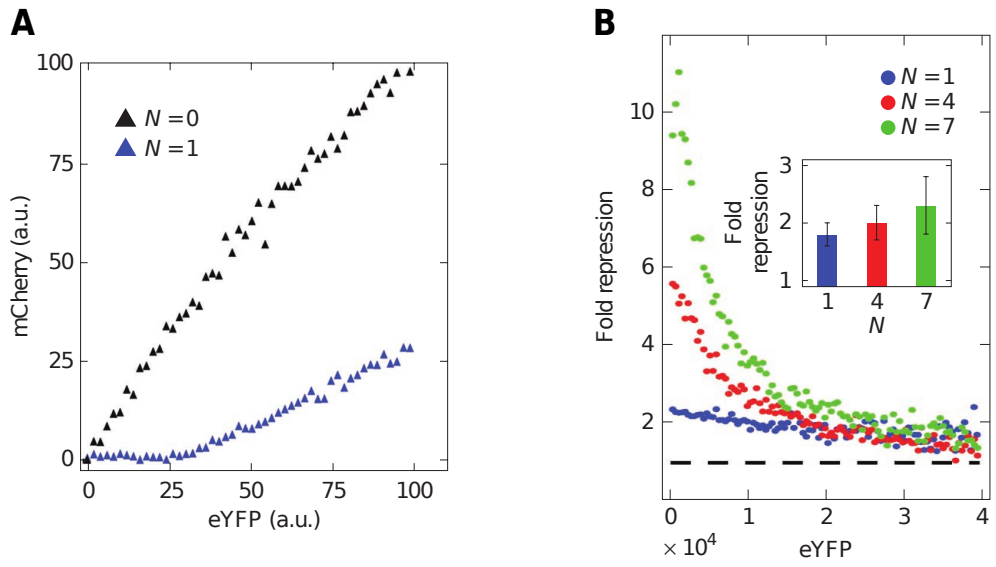


Figure 2.3: Effect of miRNA-mediated regulation on protein abundance. miRNA regulation induces a nonlinear relationship between mRNA and protein abundance. (A) Both mCherry and eYFP are fluorescent reporter proteins. The mCherry 3'UTR was engineered to contain N target sites. Individual cells were arranged depending on eYFP expression level (a.u.: arbitrary unit). In absence of target sites mCherry protein is proportionally expressed to eYFP. With $N = 1$, expression of mCherry does not rise until a threshold level is exceeded. (B) Depending on N, target repression may be up to 10-fold at low expression levels. Inset shows average fold repression for different N. Both figures are taken from [46].

developed a biochemical model that could explain both modes, i.e. the switch-like as well as the moderate control by miRNAs. The former mode is active at low target abundance. Then protein production is repressed strongly up to tenfold and more depending on the number of miRNA target sites per transcript. Interestingly, Hafner et al. determined experimentally miRNA-target interactions transcriptome-wide and found predominantly transcripts exhibiting low and medium expression levels to be involved [9]. Repression gets moderate - about twofold - when the amount of mRNA reaches a critical concentration leading to saturation of the miRNA pool. Hence, at high target levels miRNAs perform fine-tuning of gene expression. These results are further supported by a study of Arvey et al. who describe a significant anti-correlation between downregulation by miRNAs and target abundance [47]. At high target levels many molecules are competing for a

limited small number of miRNAs and thus dilute miRNA activity. Concluding, the strength of repression depends strongly on the relative amounts of the miRNA and its targets. Importantly, Mukherji et al. noticed that strength of repression varied significantly between identically prepared cells. Probably, cell-to-cell differences explain in part why studies averaging over multiple cells (populations) such as [44] and [45] measured only moderate responses.

2.3 The miRNA metabolism in animals

2.3.1 Molecular mechanisms of miRNA biogenesis and decay

Transcription of miRNA genes is typically conducted by RNA Polymerase II. Just as with mRNAs, the transcripts are capped at the 5' end and polyadenylated at the 3' end. miRNA genes are organized in various forms on the genome causing different types of primary miRNA transcripts (pri-miRNAs). While some miRNAs originate from individual transcripts a greater fraction of miRNA genes are part of miRNA clusters yielding *polycistronic* transcripts [48]. Moreover, another appreciable fraction of miRNA genes ($\sim$ 50 in vertebrates) is located within loci of protein-coding and non-coding genes, predominantly in intronic regions. These miRNAs are either co-expressed with their host gene or they are independently transcribed. According to a study of Monteys et al. approximately 35% of intronic miRNAs in vertebrates have own promoters [49].

The canonical pathway of miRNA maturation involves two processing steps catalyzed by enzymes of the RNase III protein family, Drosha and Dicer. Within the nucleus Drosha excises with aid of the cofactor DGCR8, a double-stranded RBP, the *stem-loop* or *hairpin* structure from the remainder of the pri-miRNA to create the $\sim$ 70 nt long pre-miRNA intermediate. In rare cases pre-miRNAs are obtained from an alternative miRNA biogenesis pathway that bypasses Drosha-processing via the generation of so-called *mirtrons*. Mirtrons are very short introns that exhibit a hairpin structure like pre-miRNAs and enter the miRNA metabolism after they have been spliced out from the pre-mRNA [50]. The pre-miRNA is exported by the nuclear transport receptor Exportin-5 to the cytoplasm where Dicer supported by the transactivation-responsive RBP (TRBP) catalyzes the removal of the terminal loop from the hairpin structure resulting in the mature miRNA duplex of $\sim$ 22 nt length. Each of the two catalyzed cuts determines one end of the final miRNA sequence. Carthew et al. point to the necessary high precision of the processing machinery to ensure the specificity of the produced miRNA [51]. Nevertheless, several recent studies suggest that a single miRNA locus may lead to multiple variants of one miRNA species that differ in length and sequence. These

variants are called *isomiRs*. The heterogeneity can arise from imprecise cleavage but also from post-processing of the miRNA sequence by factors such as Exoribonucleases, Nucleotidyl transferases or RNA editing. In particular, the 3' end of the miRNA is subject to modifications. But variations of both the 5' end and the internal sequence have been observed as well. Furthermore, expression analysis revealed isomiR biogenesis can be cell type specific [52].

One strand of the duplex associates with a member of the Argonaute (Ago) protein family while the other strand - denoted miRNA* or passenger miRNA - disappears. As association to Ago and cleavage by Dicer are physically coupled, the mature miRNA duplex does not persist long in the cytoplasm [51]. The selection which strand binds to Ago depends on the thermodynamic stability of the duplex's ends. Typically, the strand that is less stably base-paired at the 5' end is selected [53]. But exceptions to this rule, i.e. miRNA* is associated with the Ago, have been detected as well in considerable number. Ago is the central component of the multiprotein-containing RNA-induced-silencing complex (RISC). Beside miRNA-bound Ago, the GW182 protein is crucial for gene silencing. The miRNA serves as specificity factor that guides the RISC to the transcript to be regulated [51]. Regulation by the RISC leads either to translational repression or to mRNA degradation, see section 2.4.

Unlike biogenesis of miRNAs that is well understood the exploration of miRNA stability and decay is in its beginnings. The knowledge is still very patchy and thus it is not yet possible to create a comprehensive model as it does exist for the biogenesis. The delayed onset of investigative efforts in this field was partly due to the initial perception of miRNAs to be highly stable molecules with half-lives of many hours or even days [54]. Though recent studies reported on accelerated decay of miRNAs and great diversity of miRNA degrading enzymes across phylogeny [55]. These findings indicated that the importance of miRNA turnover most likely had been underestimated. For instance, Krol et al. analyzed the miRNA turnover in retinal neurons and measured half-lives of $\sim$ 1 hour [56]. Moreover, given that miRNAs are involved in developmental transitions, it seems obvious that the spectrum of miRNA half-lives includes rapid turnover times as well [54].

Concerning the molecular mechanisms, miRNA degradation is carried out by exoribonucleases [55],[57]. In general, unprotected ends leave miRNAs as other RNAs vulnerable to degradation by nucleases. Various 5'-to-3' and 3'-to-5' exoribonucleases have been identified that degrade miRNAs by removing terminal nt from either the 5' end or the 3' end, respectively. The nucleases known to be involved in miRNA turnover in human are XRN1, RRP41 and PNPase$^{\text{old-35}}$ [55].

2.3.2 Regulation of miRNA metabolism and function

miRNAs are important for the regulation of numerous cellular processes and contribute greatly to context-specificity of protein expression. Many of them were shown to be expressed depending on tissue or developmental stage. Further, changes in miRNA expression profiles are associated with many diseases [58]. Here, the various possibilities how miRNA abundance and function are regulated are shortly introduced by means of examples.

Regulation of miRNA gene transcription Intergenetic miRNAs but also in part intragenic miRNAs are autonomously expressed and thus have own promoter regions. These regions are equipped with typical promoter elements such as CpG islands, TATA box sequences, initiation elements and histone modifications suggesting transcriptional activation of miRNA genes to be similarly regulated to that of protein-coding genes [54]. Importantly, many examples of miRNAs being involved in their own (de-)activation via autoregulatory *feedback loops* have been discovered. For instance, Kim et al. found a negative feedback loop constituted by miR-133b and Pitx3 that regulates maturation and function of midbrain dopaminergic neurons. Pitx3 activates the transcription of miR-133b and the miRNA suppresses the expression of Pitx3 post-transcriptionally [59].

Post-transcriptional regulation of miRNA expression During maturation the nascent miRNA is processed by Drosha and Dicer. Both proteins as well as their respective interaction partners, DGCR8 or TRBP, are subject to regulation that in turn can affect miRNA expression. For instance, the complex formation with DGCR8 stabilizes Drosha. On the other hand, Drosha may induce degradation of DGCR8 by cleaving hairpins present in the DGCR8 mRNA [54], probably because high concentration of DGCR8 was shown to inhibit Drosha processing of pri-miRNAs [60]. Recently, the terminal loop of the hairpin structure was found to be used as interaction site by different RBPs during Drosha processing. By rearranging the structure, uridylation or cleavage RBPs can enhance or inhibit the progress of miRNA maturation [61]. In matters of Dicer, a highly influential negative feedback loop was identified to exist between miRNA let-7 and Dicer. The former has been shown to regulate Dicer post-transcriptionally [44],[62]. Consequently, let-7 may affect globally the maturation of miRNAs within the cell [54].

The excision of the miRNA duplex from the pri-miRNA transcript determines the terminal nt of the miRNA strands. As noted above, the thermodynamic stability of the 5' ends is decisive for which of the two strands is preferentially loaded into the RISC. Therefore, cleavage heterogeneity caused by Drosha or Dicer in-

fluences strand selection. Further, an altered 5' end affects the function of the miRNA as the 5' end is crucial for the target recognition [54], see section 3.1.1.

Regulation of miRNA function The mature miRNA is incorporated into the RISC and guides the gene-silencing complex to its targets. Control at this level concerns miRNA function or stability. miRNA function can be affected in a variety of ways. Crucial are the two core protein components of the RISC, Ago and GW182. Regulation of these has effect on miRNA function. Additionally, each of them is associated with numerous other proteins that contribute to miRNA-mediated control increasing the pool of targets for controlling the miRNA pathway [54].

Both Ago and GW182 have paralogues in many organisms. In vertebrates Ago is present in four variants. Regarding target selection as well as miRNA selection the four Ago proteins in human do not differ from each other [9]. But the power to prevent protein synthesis varies between them [54]. Further, as noted in section 2.2.3, miRNA abundance and RISC formation are limited by the concentration of the Ago proteins. Hence, the expression levels of Ago proteins might control the extent of miRNA regulation. Several regulatory mechanism have been identified that control Ago expression such as the heat shock protein 90 that is important for stabilization. Further, a couple of protein modifications with modulating effects are known [54]. GW182 has three paralogues in mammals, the trinucleotid-repeat containing proteins TNRC6A, TNRC6B and TNRC6C. Unlike Ago, the investigation of the regulation of GW182 proteins is in its infancy and thus it is less known about the their regulatory impact on miRNA function [54].

RBPs are considered to play a major role in post-transcriptional regulation [63]. They are involved in maturation, localization, translation, stability and degradation of mRNAs. Opposed to miRNA-mRNA associations that depend on the presence of sequence signals on the mRNA, RBPs recognize particular secondary structural elements. One of the first examples demonstrating the interplay between the two classes of regulators involves miR-122 and the RBP HuR [64]. miR-122 inhibits translation of the CAT1 mRNA under normal conditions in human hepatoma cells. Upon stress induction HuR binds to the 3'UTR of CAT1 and thereby dissolves miRNA-mediated repression. The repressed mRNAs are then redirected to polysomes for active translation. Interaction between HuR and miRNAs is potentially widespread. By experimental studies hundreds of HuR binding sites were found adjacent to or overlapping with miRNA target sites [65].

In case of HuR and miR-122 the RBP counteracts miRNA regulation but examples of cooperation between both types of regulators have been observed as well.

Kedde et al. showed that miRNA-mediated repression of the tumor suppressor p27 is enabled by the RBP PUM1 [66]. In quiescent cells, both protein p27 and its suppressors miR-221 and miR-222 are expressed at high level. The structural conformation of the p27 3'UTR prevents miRNA binding. PUM1 is upregulated in response to growth factor stimulation. By binding to 3'UTR of p27, PUM1 induces structural reformations that unfold the target sites of miR-221 and miR-222. Subsequent repression of p27 by the miRNAs leads to rapid entry to the cell cycle. As for HuR, global analyses revealed RBP binding motifs to be enriched in 3'UTRs with miRNA target sites and vice versa [63]. Further examples of miRNA-RBP interactions are known [54],[63],[65]. As the numbers of different miRNAs and RBPs respectively reach several hundreds in animals, cross-talk between the two classes of regulators is supposed to be a general mechanism [63].

Regulation of miRNA degradation Recent studies revealed that many miRNAs are subject to accelerated or regulated miRNA turnover [54],[55]. For instance, miR-29a and miR-29b originate from the same locus and thus are co-transcribed but their mature forms showed different expression levels during cell cycle progression of HeLa cells. Hwang et al. observed that miR-29a is constitutively expressed in all phases of the cell cycle whereas miR-29b is enriched in mitotic cells but depleted in the remaining phases [67]. The regulation here acts on the mature miR-29b as transfected mature miR-29b showed a similar expression pattern with accumulation in mitotic cells. Compared to miR-29a, miR-29b experiences an accelerated decay triggered by the cell cycle. Further analyses revealed that centrally arranged uracils in the mature miRNA sequence are involved in the fast degradation.

Krol and colleagues found that several miRNAs in mouse retina are light-regulated [56]. Shifting the mice from light to dark led to a reduction of the levels of miR-204 and miR-211 as well as of miRNAs from the cluster miR-183/96/182. But also miRNAs insensitive to light showed rapid turnover in neurons. They observed similar turnover of miRNAs in hippocampal slices and in neurons differentiated from mouse embryonic stem cells suggesting rapid miRNA turnover to be a general feature of neurons. Importantly, miRNAs in non-differentiated neurons were not subject to fast decay. Further, turnover in neurons is supposed to be activity-dependent as blocking of action potentials prevented fast decay. Both studies are examples of accelerated decay of miRNAs induced by physiological triggers.

The stability of mature miRNAs is controlled by cis-regulatory elements, cis-acting modifications (i.e. alterations of the miRNA sequence), protein complex

formation and exposure to nucleases [57]. It has been shown that addition of adenine residues to the 3' end conferred protection to the miRNA molecule [57]. On the other hand, Baccarini and colleagues observed that attachment of uracils to the 3' end triggered decay of miRNAs [68]. Importantly, cis-acting elements causal for degradation have been identified along the entire miRNA [55]. The incorporation into the RISC by binding to the Ago protein stabilizes the miRNA by protecting it from exoribonucleases. Observations suggest that miRNAs compete among each other for association with Ago. Therefore, saturation of Ago-binding capacity is assumed to limit function and stability of miRNAs [57]. Furthermore, Chatterjee et al. found that in absence of targets miRNAs dissociate from the RISC and become degraded [69]. Importantly, this result demonstrates that miRNA stability depends also on the target pointing to a bidirectional regulatory relationship between miRNA and mRNA.

2.4 Regulation of Translation by miRNAs in animals

2.4.1 Molecular mechanisms involved in miRNA regulation

Translation is a three-step process including initiation, elongation and termination. The 80S ribosome complex is assembled at the initiation stage and positioned at the translation start site of the mRNA. During elongation, the ribosome moves along the coding sequence in a 5' $\rightarrow$ 3' direction and synthesizes the peptide chain. Release of the newly produced protein and removal of the ribosome complex terminates the translation [70].

Typically, mRNAs are cap-dependently initiated, that is, the mRNA needs to be equipped with a cap structure at the 5' end. At the 3' end, the mRNA is prolonged by a sequence of adenines, called poly(A) tail. Both 5' cap and poly(A) tail have several functions, but particularly each of them contributes to the stability of the mRNA. Deadenylation and decapping of the mRNA, i.e. removing both the poly(A) tail and the cap, leads to decay of the mRNA by exoribonucleases [71]. Proteins that associate with the cap and the poly(A) tail during initiation interact with each other and give rise to a circular structure of the mRNA. The eukaryotic translation-initiation factor (eIF4F), which consists of the cap-binding protein eIF4E, an helicase eIF4A and the scaffold protein eIF4G, recognizes the cap structure. Interaction between eIF4G and the poly(A)-binding protein (PABP) promotes circularization, see figure 2.4. mRNAs in closed loop configuration are efficiently translated and protected from enzymatic degradation [43].

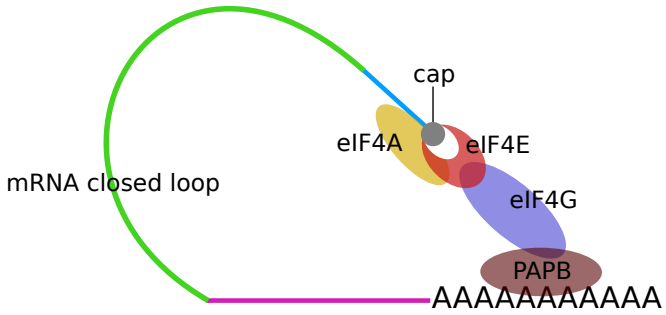


Figure 2.4: Circular mRNA. eIF4F interacts through its cap-binding subunit eIF4E with the 5' cap structure of the mRNA. eIF4A unwinds secondary structures in the 5'-untranslated region (5'UTR) enabling the translation machinery to scan the 5'UTR sequence towards the initiation codon. PABP binds to the poly(A) tail and complexes with the scaffold protein eIF4G to close the mRNA loop. Coloring of the mRNA regions: 3'UTR (magenta), coding region (green) and 5'UTR (blue). This figure is based on a figure from [43].

2.4.2 Current model of gene silencing through miRNAs

Cases of translational activation by miRNAs have been reported. In quiescent cells certain miRNAs increase translation of mRNAs [72]. Activation by miRNAs is mediated either by relief of repression or by direct activation. But currently translational activation appears to be restricted on specific G0 states of the cell cycle. Therefore, unlike silencing, activation is not expected to be a general mechanism of the miRNA pathway [71],[73].

miRNA lin-4 inhibited translation of lin-14 mRNA with no discernible effects on mRNA abundance [3]. As ribosomes were found attached to the regulated lin-14 mRNA molecules, it was concluded that suppression by lin-4 occurred after translation had been initiated, i.e. at the elongation stage [74]. Since then a variety of observations were made. Some of them agree with this early model but others are not covered by it or are even not compatible with it. Meanwhile, Bagga et al. found in a re-evaluation that lin-14 mRNA level is significantly decreased by lin-4-mediated regulation [75]. According to the current understanding, miRNAs bind to the target at the stage of translation initiation since several studies suggest that miRNA regulation prevents ribosome recruitment [43],[76]. Additionally, it has been demonstrated that the RISC binds to the 5' cap structure and impairs translation initiation. Consistently, alternatively translated mRNAs that lack the 5' cap structure are refractory to miRNA-mediated silencing.

The initial view that miRNAs inhibit translation but do not affect mRNA abundance had to be extended. Although this first notion was supported by subsequent studies, others reported strong correlation between decreases in protein level and mRNA level in response to miRNA silencing [77]. Decreased mRNA levels were due to miRNA-mediated deadenylation, decapping and ensuing degradation of the mRNAs. Moreover, Bhattacharyya et al. found that miRNA-bound mRNAs may be translationally reactivated upon exposing cells to specific stress [78]. These findings raised the question how the molecular events (translational repression, mRNA deadenylation and mRNA decay) relate to each other. Is the perceived translational repression the result of mRNA decay or does destruction of the mRNA follow on translational inhibition [79]?

Two time-course analyses provided revealing insights. Djuranovic et al. monitored protein expression and mRNA level of miRNA targets in fruit fly cells at different time points [80]. Repression of translation was consistently observed to occur already before deadenylation and degradation. Further, they found that repression is independent of deadenylation. In a parallel study Bazzini et al. studied miRNA effects in zebrafish embryos [81]. miR-430 was known to clear endogenous maternal mRNA through deadenylation and ensued degradation. But whether translational repression preceded deadenylation of mRNA was unknown. Consistent with the results of [80] they could demonstrate that miR-430 represses initiation of translation before initiating deadenylation and decay.

In mammalians the impact of translational repression without mRNA decay was rated low by some global studies. For instance, both Baek et al. and Selbach et al. detected significant correlation between mRNA and protein levels of miRNA targets [44],[45]. Further, Guo et al. concluded that miRNAs predominantly act to decrease target mRNA levels [82]. Hu and Coller conjectured these assessments have to be attributed to the conduction of the experiments, in particular to the time points at which miRNA effects were studied in these works [79]. The earliest time point analyzed in the study of Guo et al. was 12 hours after transfection of the miRNAs. In contrast, the earliest time point considered in the fly and fish studies was 2 hours. Selbach et al. measured two time points, 8 and 32 hours after transfection [44]. Interestingly, at 8 hours the correlation between mRNA and protein levels was weak for many genes but it became strong at 32 hours. Possibly, by considering late time points, the mammalian studies missed the phase of translational repression and measured primarily the subsequent mRNA decay.

Based on these findings, Fabian and Sonenberg proposed a temporal model of miRNA-mediated repression [77]: miRNAs act through a two-step mode of repression. First, translation is inhibited independent of the adenylation status of the mRNA, see figure 2.5. The molecular mechanisms underlying repression are

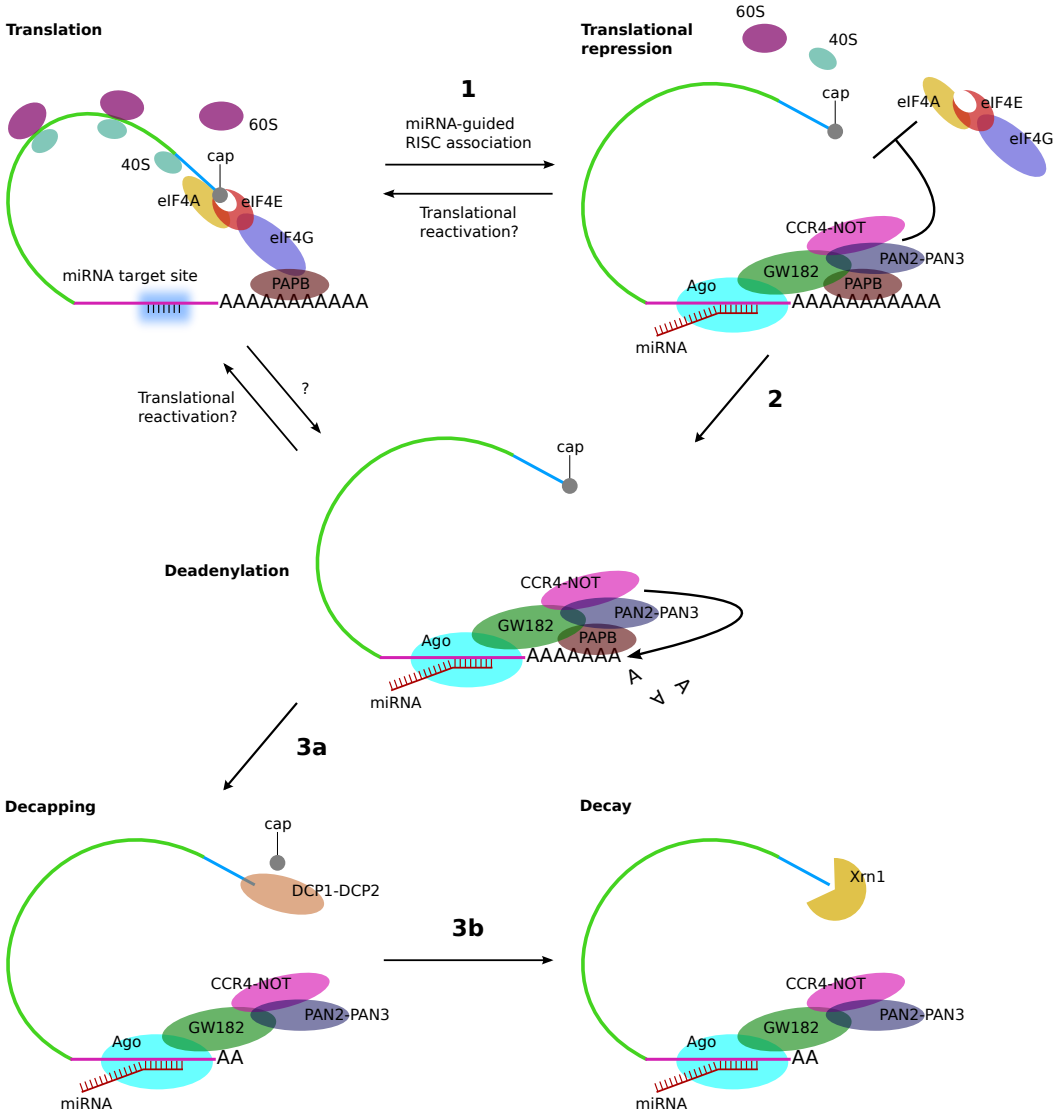


Figure 2.5: Temporal order of miRNA-mediated gene silencing. The Ago-bound miRNA guides the RISC to the target mRNA (1). Various models of RISC-induced inhibition of translation are currently discussed. Interaction of GW182 with PABP interferes with the closed loop configuration of the mRNA and prevents ribosome recruitment. Further, the CCR4-NOT deadenylase complex is necessary for repression. Following repression, mRNA decay is initiated through deadenylation by CCR4-NOT and PAN2-PAN3 (2). Decapping by DCP1-DCP2 (3a) and degradation by 5'-3' exoribonuclease Xrn1 (3b) finishes mRNA decay. Progressions denoted by question marks can not be ruled out presently.

not fully characterized yet. It is assumed that the GW182 interaction with PABP competes against the eIF4G-PABP association and thus prevents circularization of the mRNA. Further, there is evidence that the CCR4-NOT deadenylase complex is involved in repression independent of its deadenylase activity. Subsequently, suppressed mRNAs enter the cellular 5'-3' mRNA decay pathway [43]. Deadenylation by CCR4-NOT and PAN2-PAN3 deadenylases initiates the mRNA decay. Following poly(A) tail removal, the 5' cap structure is dissociated by the DCP1-DCP2 decapping complex and the unprotected mRNA molecule is degraded in 5' $\rightarrow$ 3' direction by the exoribonuclease Xrn1. A couple of questions remain unanswered, e.g. if deadenylated mRNAs may be reactivated or if repression needs to precede degradation.

3 Analysis of miRNA seed-based target recognition

3.1 Background: Target recognition by miRNAs in animals

"At the heart of miRNA-mediated regulation lies the species-specific interaction of the miRNA and the mRNA" [55]. Within the RISC, miRNAs act as adapters that confer target-specificity on the gene silencing machinery. Main object of this thesis was to study how miRNAs carry out their adapter functionality, that is, it was questioned how do miRNAs recognize target mRNAs. The following describes the general principles underlying miRNA target recognition in animals and introduces biochemical methods for uncovering miRNA targets in a large scale.

3.1.1 The miRNA 5' region is crucial for target recognition

By formation of a RNA-RNA duplex between the miRNA and a complementary site on the target, the target mRNA is selected for downregulation through RISC. Unlike plants, where the majority of targets bear sites of extensive complementarity to the miRNA, broad complementarity is unusual in animals [7]. The first discovered miRNA target in animals, lin-14, contained seven conserved sites in its 3'UTR, each harboring an almost exact copy of a 9 nt sequence motif. This "core element" was perfectly complementary to the 5' region of lin-4 miRNA. Base-pairing between the remainder of the miRNA and the target was fragmentary and varied between the target sites [3]. Subsequent observations by others underpinned the particular relevance of the miRNA 5' region for target recognition. *K box* and *Brd box* were known 3'UTR sequence motifs in Drosophila that mediate post-transcriptional regulation. Lai found that 5' regions of Drosophila miRNAs were perfectly complementary to these motifs [83]. Sequence comparison of nematode miRNAs revealed that the 5' end of miRNAs is more conserved than the 3' end [84]. This finding also brought forth the concept of miRNA *families*. miRNAs independent of their genomic origin were sorted to miRNA families if they shared sequence identity minimum in the 5' region. Lewis et al. from the Bartel laboratory finally documented the importance of the 5' region for target

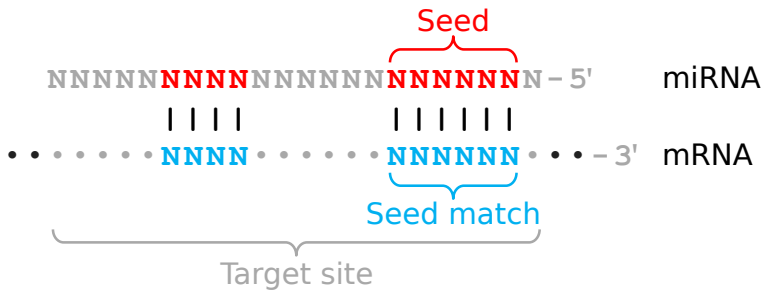


Figure 3.1: SCHEMATIC REPRESENTATION OF A MIRNA TARGET SITE. Typically, the miRNA seed located at the 5' end of the miRNA matches perfectly with the target mRNA. The seed-complementary site on the target is referred to as the seed match. Unlike the seed, the 3' portion of the miRNA base-pairs less regularly with the target. The region on the target comprising the entire interaction between miRNA and target is called target site.

recognition by computational analysis [85]. In this article also some basic terms and concepts regarding miRNA target recognition were introduced, see figure 3.1. The perfectly pairing sub-segment of the miRNA was termed the miRNA *seed* and the corresponding complement on the 3'UTR was referred to as the *seed match*. The latter in turn is part of the *target site* on the 3'UTR that may involve additional but much less regular base-pairing between miRNA and 3'UTR. Hence, computationally, miRNA target sites can be uncovered by seeking for seed matches.

Lewis and colleagues aimed for verifying the relevance of the 5' region for target recognition without consideration of the few known targets. Thus, all sub-segments of the miRNA were assessed; each segment was used as seed for the search for target sites on human 3'UTRs. Only if a miRNA-3'UTR interaction was conserved, that is, orthologous 3'UTRs in mouse and rat contained minimum one target site for the same miRNA, the interaction was counted. Conservation indicates preserved biological function. Therefore, Lewis et al. reasoned that conserved miRNA-target interactions are very likely true interactions. It is worth mentioning that they did not require a target site to reside at orthologous positions in the orthologous 3'UTRs to be considered conserved but only to be present (somewhere) in these 3'UTRs.

The 7 nt long segment ranging from position 2 to 8 relative to the 5' end of the miRNA exhibited the highest signal-to-noise ratio, i.e. the number of detected targets exceeded the number of targets expected by chance most. Targets found for randomly permuted ("control") miRNAs constituted the expected fraction or

noise. This result confirmed the previous observations that the 5' located seed region is most important for target identification. Other segment lengths were probed as well, but shorter segments gave rise to too low signal-to-noise ratios and longer ones retrieved less targets at comparable signal-to-noise ratio. Lewis et al. implemented these findings in the miRNA target prediction software TargetScan [85]. The first version of TargetScan predicted more than 400 mammalian genes to be conserved targets of miRNA regulation.

3.1.2 Different types of seed-complementary sites

In following years the groups of Bartel and Burge made use of the steadily growing number and the increasing quality of fully sequenced genomes to further enlighten target recognition by miRNAs. Requiring the seed match to occur at orthologous positions in five vertebrate genomes reduced the fraction of false positive predictions such that already 6 nt long seed matches, including nt 2-7, achieved modest specificity [86]. When exploring conserved positions flanking the 6mer seed match on the target, they observed that adenines were over-represented at the first position downstream of seed matches regardless of whether the miRNA starts with an uracil or not. They suspected these adenines to be involved in miRNA-target interaction, even though not necessarily through base-pairing with the miRNA. And indeed, filtering seed matches with an adenine next to the seed match improved specificity of target predictions for all miRNAs. Furthermore, the nt directly upstream of the seed match had a propensity to be base-paired to nt 8 of the miRNA. Also here, focusing on seed matches extending to position 8 increased specificity. Requiring both the matching to miRNA nt 8, denoted *m8*, and the A at position 1 in the target, denoted *A1*, yielded the highest signal-to-noise ratio. Noteworthy, all filterings, in particular requiring simultaneously m8 and A1, were accompanied with substantial losses in sensitivity. Without filtering 13,044 regulatory interactions involving 5,300 human genes could be identified based on an alignment of four mammalian genomes.

Further refinements of the method led to the discovery of the *offset 6mer* site type matching miRNA positions 3-8 by Friedman et al. [87]. First, the noise estimation was revised and has been correspondingly renamed to background estimation. Additionally, a quantitative approach based on phylogenetic trees was established to measure the conservation of seed matches. Trees were constructed based on a 23-way alignment of vertebrate 3'UTRs. Summing the lengths of the branches connecting the subset of species having a seed match perfectly aligned produced a branch-length score for the seed match. This allowed to analyze signal-to-background ratios at varying branch-length cutoffs each representing a different

Figure 3.2: SET OF CLASSICAL SEED TYPES. The core seed region includes nt 2-7. Three of the seed types base-pair additionally with nt 8 of the miRNA. Albeit the adenine opposed to miRNA residue 1 is not necessarily involved in base-pairing, its presence increases specificity of target site prediction.

level of conservation.

Other 6 nt long site types beside 6mer and offset 6mer did not show appreciable signal-to-background ratios. Therefore, Friedman et al. focused in their analysis on following site types: 6mer and offset 6mer, 7mer-A1 and 7mer-m8 and the 8mer, that is the combination of the two 7mer types. Henceforth, within this thesis the term *seed type* is used instead of site type, see figure 3.2. Figure 3.3(A) displays the number of conserved sites for each seed type for two branch-length cutoffs. With regard to target site prediction requiring strong conservation caused high specificity, represented by the high signal-to-background ratios. Particularly, seed matches related to the two 7mer and 8mer seed types are subject to purifying selection. On the other hand, more sites above background, reflecting sensitivity, are detected at the less stringent cutoff. Seed type 7mer-m8 is most important in terms of number of sites under selection (conserved sites above background), while the proportion of seed matches being selectively maintained (represented by the signal-to-background ratio) is highest for the 8mer. Regardless of the cutoff, ranking the seed types according to the signal-to-background ratios resulted always in the same hierarchy, with 8mer > 7mer-m8 > 7mer-A1 > 6mer > offset 6mer. Moreover, Friedman et al. measured the impact of seed types on transcript stability and found that the extent of transcript destabilization upon miRNA transfection correlated well with the fraction of conserved sites, see figure 3.3(B). This observation suggested a strong relationship between selective maintenance of seed types and their efficacy. Finally, at the most sensitive cutoff minimum 44,000 target sites on 9,600 genes were identified to be conserved above background when merging

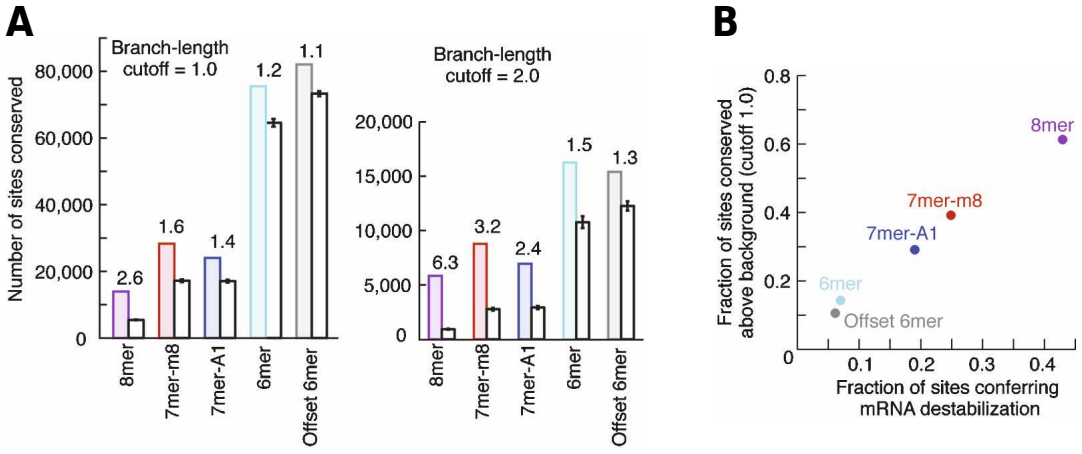


Figure 3.3: Characteristics of classical seed types. (A) Conservation of seed types at high sensitivity (cutoff = 1.0) and high specificity (cutoff = 2.0) (solid bars). Open bars indicate the background estimation. The signal-to-background ratios are denoted above the bar pairs. (B) Correlation of seed type conservation and mRNA destabilization. Both figures are taken from [87]

the seed matches of the five seed types.

3.1.3 Assessing the seed rule

Constant enhancement of the analytical methods increased both sensitivity and specificity of target site identification. The first version of TargetScan predicted 400 genes to be putative miRNA targets. Current estimates suggest that about half of human protein-coding genes are post-transcriptionally regulated by miRNAs [65]. Perfect Watson-Crick pairing between the seed of the miRNA and the target initiates miRNA-target interaction and activates regulation through the RISC. The groups of Bartel and Burge identified five types of miRNA seeds, see figure 3.2, for each of which they found thousands of conserved complementary sites (much more than expected by chance) in the human transcriptome. With the discovery of the *classical* seed types the groups of Bartel and Burge bestowed fundamental insights on the miRNA research community. Dozens of target site prediction tools that have been developed meanwhile make use of these findings, see section 4.1.

Nevertheless exceptions to the classical seed rules have been reported. For instance, Baek et al. and Selbach et al. measured the effect of miRNA-mediated control on protein output. Although the majority of downregulated genes in-

cluded perfect seed matches, a substantial fraction of direct targets were devoid of perfect complementary sites [44],[45]. Vella et al. analyzed the regulation of lin-41 by miRNA let-7 in *C. elegans* and detected only imperfect seed matches containing bulges or G:U wobble pairs [88]. Applying the approach used to reveal the classical seed types was less helpful to decipher rules of non-perfect seed matching: Friedman et al. found at most small fractions of imperfect sites to contain either a single mismatch, a G:U wobble, a bulged nt within the site or within the miRNA to be under selection [87]. In another study the group of Bartel presented a class of non-classical target sites, called *centered sites* [89]. Centered sites perfectly base-pair to nt 4-14 or 5-15 of the miRNA without substantial pairing to the seed region of the miRNA. But this class of sites is present only in a marginal fraction of target sites [90]. Beside missing functional targets, classical seed types comprise false positives as well. For instance, Didiano et al. could show that perfect seed matches are not generally a reliable predictor for miRNA-target interactions [91]. Consequently, perfect seed matches are neither necessary nor sufficient for all functional miRNA-target interactions [92].

Recently, new biochemical methods (HITS-CLIP, PAR-CLIP) have emerged that allow determination of miRNA target sites in a large scale, see section 3.1.7. Genome-wide maps of RISC-target interactions were revealed by immunoprecipitation of the Ago protein and subsequent sequencing of the isolated and coimmunoprecipitated mRNA fragments. The availability of genome-wide target site maps allowed quantitative assessment of the seed rules in terms of sensitivity and specificity. For instance, Chi et al. found 73% of target sites decoded by HITS-CLIP to be equipped with classical seed matches [92]. As Bartel and colleagues required sites to be selectively maintained, seed types enriched among less conserved sites were missed or low-weighted by their approach. Therefore, experimentally determined target sites without perfect seed matches (orphans) likely include imperfect matches or less conserved seed types. By focusing on orphans, Chi et al. discovered an alternative mode of miRNA target recognition based on *bulged sites* [92]. Here, target recognition proceeds in two consecutive phases. First, in a transitional nucleation phase miRNA positions 2-6 base-pair to an uninterrupted complementary site of 5 nt on the target. Then, a more stable duplex is obtained that includes minimum miRNA residues 2-7, with the target nt that originally base-paired with residue 6 bulging out from the pairing. Residue 6, termed the *pivot* residue, plays a special role here as it pairs with two different nt during target recognition. Bulged sites are conserved and comprised about half of the orphan sites.

3.1.4 Seed type definition based on CLIP data

We used genome-wide miRNA-mRNA interaction maps to define functional miRNA target sites [1]. Seed matches located in mRNA regions involved in interaction with Ago were classified functional and matches beyond these regions were tagged non-functional. Based on the ratio of functional-to-non-functional sites, a minimal and sufficient set of seed types was defined. As the method has been used in this thesis, it is briefly explained here: Only seed matches starting at positions 1, 2 or 3, subsequently denoted α, β and γ, relative to the 5' end of the miRNA were taken into account for the deduction of seed types. A *seed match type* is defined by its length l and its start position. A seed type may comprise several seed match types of consecutive lengths sharing the start position. For each of the three start positions following recursive procedure has been applied. In case the proportions of functional and non-functional sites of seed match type a with $l_a = k$ varied significantly from that of the combined seed match types $a+$ with $l_{a+} > k$, a was defined as seed type and the procedure was repeated for the next seed type length $k + 1$. Otherwise the method terminated and the seed match types with $l_a = k$ and $l_{a+} > k$ were merged into one seed type.

Unlike Bartel and colleagues, we had not to demand selective maintenance as proof of authenticity as our set of functional sites was experimentally verified. In fact, we could study conservation as a property of seed matches. Considering non-conserved sites extended the pool of examined target sites remarkably consistent with previous studies suggesting non-conserved targeting to be widespread [45],[93]. Although the majority of functional sites matched classical seed types, we established a new class comprising less conserved sites base-pairing with miRNA positions 1-6 that substantially contributed to sensitivity.

3.1.5 miRNA regulation through the coding region

Until recently investigation of miRNA function was almost exclusively confined to target sites located in the 3'UTR. Coding region or open reading frame (ORF) and 5'-untranslated region (5'UTR) received less attention, although experimental and computational analyses suggested involvement in particular of the coding region in miRNA targeting [45],[86],[94]. The 5'UTR as well may host effective target sites as artificial insertion of let-7 seed-complementary sites in the 5'UTR had been demonstrated [95]. Furthermore, in plants miRNAs interact primarily through sites residing within the coding region [96]. A couple of reasons may explain the bias to the 3'UTR: Unlike the ORF that was clearly associated with a biological function, the functional role of the 3'UTR was still awaiting its elucidation [97].

Additionally, it was known that the 3'UTR interacts with protein complexes performing post-transcriptional regulation comprising translational control, subcellular localization and mRNA stability. Further, the average 3'UTR length appeared to correlate with organismal complexity [98]. Thus, there were some understandable and founded reasons advising to link miRNA-mediated regulation with the 3'UTR. But it can not be ruled out that at least in part simply the first described miRNA-target interaction, with lin-4 pairing with sites in the lin-14 3'UTR, may have directed subsequent studies [48].

Apart from arguments for the 3'UTR, there were biological considerations but also technical difficulties disfavoring the coding region: Especially mechanistic aspects argued against interaction of miRNAs with the coding region. It was suspected that the protein translation machinery interferes with RISC activity [48]. Further, considering the intensity of repression, several experimental analyses demonstrated that target sites in the coding region albeit abundant are much less effective than seed matches in the 3'UTR [9],[45],[94],[99], see figure 3.4(A). Finally, the ORF was assumed to lack necessary flexibility to accumulate miRNA target sites due to the strong conservation of the amino acid sequence [97]. In terms of biology this turned out to be a weak argument as the degeneracy of the code allows for much more information to be encoded [100]. But with regard to methodology, seeking for significant signals in the highly conserved coding region is much more challenging than in the 3'UTR [48].

Numerous research groups contributed to the current multifaceted understanding of miRNA-mediated regulation by studying the interaction between miRNAs and target sites in the 3'UTR. Like a *preferential attachment process* the number of publications dealing with 3'UTR-related miRNA action was growing faster and faster whereas studies of the coding region remained the exception, even though interim computational analyses found thousands of target sites in the coding region [86],[94]. Eventually, the great number of functional sites discovered by the CLIP methods [8],[9] and by computational approaches adapted to the coding region [62],[101] gave rise to increased interest in miRNA target sites located in the coding region.

Forman et al. were the first who presented a method to identify conserved elements in the coding region [62]. The coding region is already highly conserved due to the strong selective pressure at the protein level. This bias has to be accounted for to find sequences being additionally selectively maintained at the nt level. Seed matches of miRNAs, including miRNA let-7, were among the top scoring motifs in human coding regions. Forman et al. found that highly conserved let-7 sites in the coding region are similar effective as 3'UTR sites, see figure 3.4(B). Interestingly, they could show that the pre-miRNA processing enzyme Dicer is downregulated

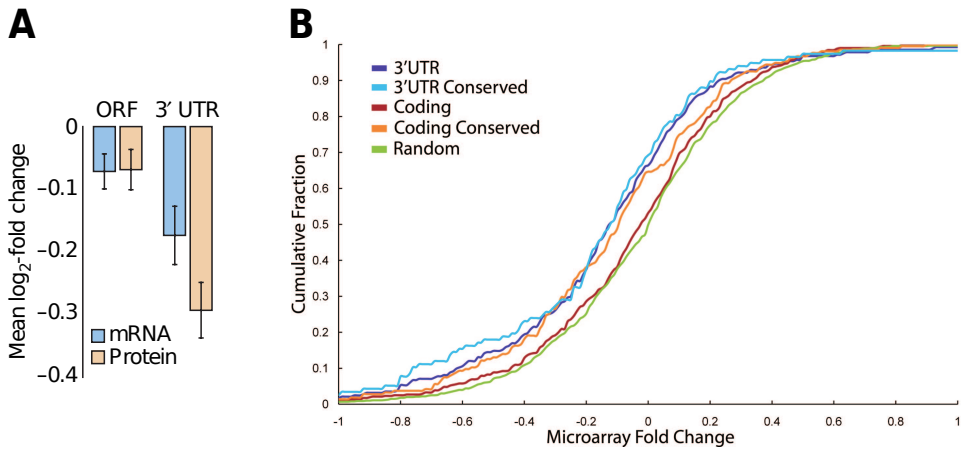


Figure 3.4: Different efficacy of ORF and 3'UTR targeting. (A) Impact of transfected miRNAs on mRNA and protein level. This figure is taken from [45]. (B) mRNA destabilization through sites in 3'UTR and coding region. mRNAs with target sites in their 3'UTR, in particular those with conserved sites, show a significant difference to a random set of genes. Regarding the coding region only conserved sites cause a significant deviation from the random set. This figure is taken from [62]

through let-7 target sites in the coding region, see section 2.3.2. Schnall et al. developed a more sensitive approach to decipher conserved sites and demonstrated that conserved targeting in the coding region is at the scale of conserved targeting in the 3'UTR [101].

It was known for 3'UTRs that multiple seed matches within a 3'UTR increase efficacy of target repression [9],[94],[102]. Fang and Rajewsky found that target sites located in the coding region enhance the effect of sites in the 3'UTR [103]. In another study Schnall et al. reported strong repression induced by multiple seed matches in the coding region to the same miRNA. The seed match copies could be attributed to a repeatedly encoded protein domain [104]. These evidences for effective targeting through the coding region motivated the development of adapted target prediction algorithms [105],[106],[107]. Marin et al. analyzed the precision of predictions obtained for different seed lengths (2-7, 2-8 and 1-8) [106]. As expected, the 7mer performed better than the 6mer, but they were surprised to find that the 7mer exceeded the 8mer in terms of precision. In this thesis an analysis of seed types in the coding region was performed, see section 3.4.1. Consistent with the finding of Marin et al., an increased importance of short seed

types could be observed in the ORF.

Concluding, the coding region should not be ignored. But why do ORF sites trigger reduced repression? According to current knowledge, it is, as initially suspected, due to the ribosome complexes that scan along the coding sequence thereby hindering the RISC to attach to target sites. An elegant experimental study by Gu et al. provided evidence for this hypothesis [99]. By inserting rarely occurring codons upstream of target sites within the coding region miRNA-mediated repression could be amplified; the rare codons induced a brief translational pause allowing the RISC to associate with the target site. Consequently, Gu et al. conjectured that functional sites in the coding region should be preceded by rare codons. Hafner et al. could nicely confirm this hypothesis using sites detected by the PAR-CLIP method [9]. They determined the codon usage around functional and non-functional seed matches and observed that there was a significant bias to rare codons in proximity of functional sites.

3.1.6 The ceRNA hypothesis: Targets compete for miRNAs

This subject would also fit into section 2.3.2 dealing with the regulation of miRNA function. But knowledge of miRNA target recognition is crucial for the understanding of this "new layer of regulation of miRNA activity" [109]. Experimental and computational analyses have shown that the intensity of target repression is dependent on the target concentration, see studies of Mukherji et al. [46] and Arvey et al. [47] outlined in section 2.2.3. An increase of target abundance diluted the initially high miRNA activity. Prior to these findings, Seitz proposed already an interesting hypothesis aiming to explain the great number of computationally predicted targets per miRNA in animals [110]. According to his *pseudotarget* hypothesis, a large proportion of miRNA targets act as inhibitors of miRNA activity. Like sponges these targets titrate miRNAs and thereby prevent them to regulate their authentic targets. In plants this principle of miRNA activity regulation was already known as *target mimicry* [111]. Poliseno et al. found that miRNA sponges play a role in the control of cancer [112]. The transcripts of the tumor suppressor gene PTEN and its pseudogene PTENP1 have closely related 3'UTR sequences that include target sites for the same set of miRNAs. By consuming PTEN-regulating miRNAs PTENP1 de-represses PTEN and enhances its tumor suppressor activity. Conversely, PTEN mRNAs act as decoy for miRNAs regulating PTENP1. Further, Poliseno and colleagues could show that genomic loss of PTENP1 is correlated with reduced PTEN expression in certain cancers. These results revealed a function for pseudogenes and added a non-coding function to coding transcripts.

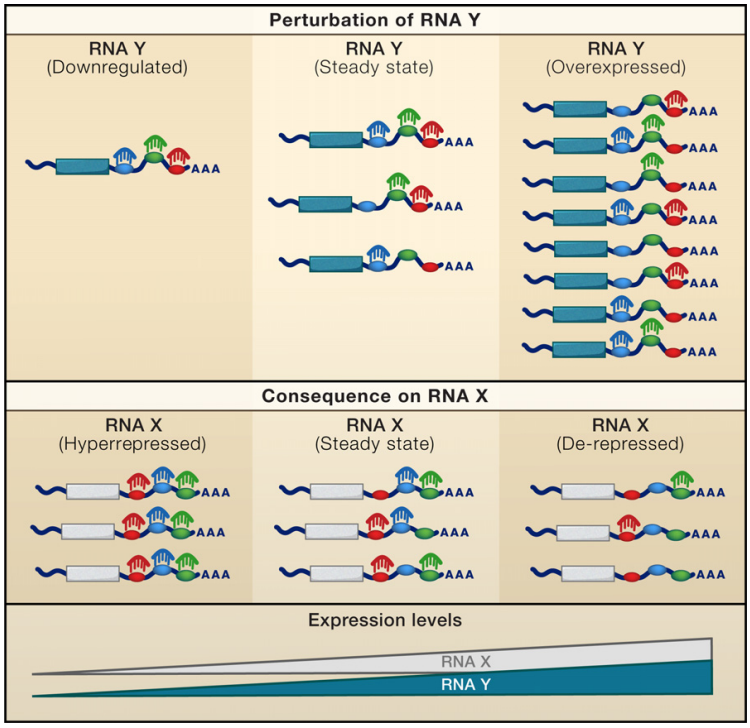


Figure 3.5: Cross-talk between ceRNAs. RNA Y and RNA X share target sites for the same set of miRNAs. Rising expression of RNA Y increases the concentration of target sites and can lead to de-repression of RNA X. Conversely, downregulation of RNA Y would lead to decreased concentration of target sites and thus to strong repression of RNA X. This figure is taken from [108].

Saleman et al. unified all these findings and considerations in a hypothesis stating that RNAs regardless whether coding or non-coding sharing miRNA target sites may affect each others' expression level by competing for miRNA binding [108]. In this model miRNA target sites constitute the "letters" of a "RNA language" that are used by competing endogenous RNAs (ceRNAs) to "talk" with each other. For completeness, the combination of target sites located on a RNA sequence is considered a "word". The more target sites ceRNAs have in common the more effective is the "cross-talk" between them, see figure 3.5. Therefore, pseudogenes and their corresponding ancestral genes are considered to be particularly suited for mutual regulation due to the high sequence similarity. Since any RNA containing at least one miRNA target site may function as a ceRNA, Saleman et al. speculated that cross-talks between ceRNAs form a large regulatory

network that includes many unknown molecular interactions. To uncover these interactions, comprehensive knowledge of number and location of the letters, i.e. the miRNA target sites, is essential. By maximizing the sensitivity in the determination of functional target sites, this work provides important groundwork for the elucidation of ceRNA networks, see section 3.4.1.

Concluding, the relationship between miRNAs and their targets is not a one-way street with miRNAs as active regulators and mRNAs as passive targets. As noted in section 2.3.2, the miRNA concentration may depend on target abundance. The ceRNA hypothesis provides another argument that miRNAs and mRNAs mutually affect each other: By acting as miRNA decoys, mRNAs and other types of RNAs may regulate the function of miRNAs without destabilizing them.

3.1.7 Experimental approaches to identify miRNA targets

Basically, experimental approaches can be subdivided in methods that are able to identify direct miRNA targets and others that can not distinguish between direct and indirect, i.e. downstream, targets. Identification of direct targets or even of the location of target sites on the target sequences is desirable but currently technically still very challenging. Therefore, established methods such as microarrays are usually used to detect miRNA effects. By measuring the change of transcript expression after overexpression or depletion of certain miRNAs, microarrays reveal both direct and indirect targets. Subsequent searching for seed matches on the sequences of affected transcripts can help to discern direct targets. A drawback is that overexpression of miRNAs may yield artifacts and repression of commonly lowly expressed miRNAs is unlikely to generate measurable effects [113].

A technique to measure protein abundance, SILAC (stable isotope labeling with amino acids in cell culture), has been used by two groups to quantify the impact of miRNA loss or overexpression on protein expression [44],[45]. Importantly, the change of protein expression represents the final result of miRNA-mediated control. But like microarray analyses, also these procedures provide both direct and indirect targets. In addition, SILAC experiments are very time-consuming and costly. For this reason, but also since mRNA and protein levels turned out to correlate well (at least for later time points, see section 2.4.2) mRNA expression studies are commonly preferred [113].

Biochemical isolation of proteins by immunoprecipitation (IP) presents a mean to uncover only direct miRNA targets. The idea is to purify the RISC together with the bound miRNA and the associated target. Easow et al. were the first who used IP for miRNA target identification [114]. They modified the Drosophila Ago1 protein such to be specifically detected by an antibody allowing pull-down

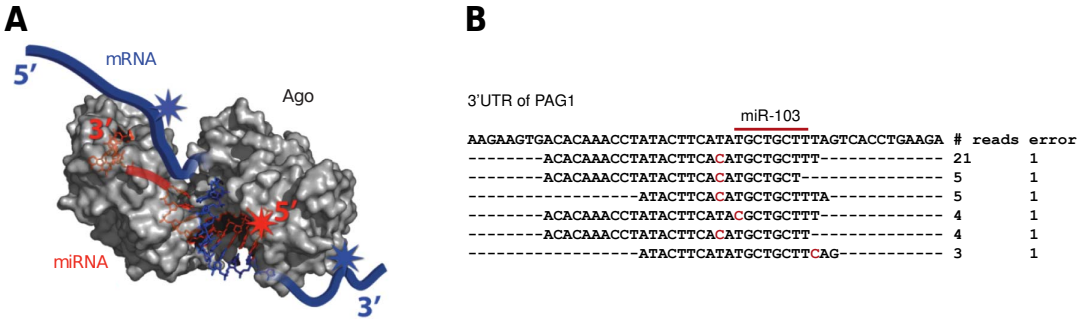


Figure 3.6: CLIP approaches. (A) Ago, miRNA and mRNA form a ternary complex allowing for simultaneous identification of the mRNA fragment and the miRNA by crosslinking and subsequent immunoprecipitation of Ago. (B) Clustered sequence reads produced by PAR-CLIP. The 41 nt long CCR is centered over the crosslinking site determined by the predominant T to C substitution. Figure A is taken from [8] and figure B is from [9].

of RISC and bound RNAs. The identity of the RISC-associated transcripts was determined with microarrays.

The HITS-CLIP (High-throughput sequencing of RNA isolated by crosslinking and immunoprecipitation) protocol represents a milestone in the development of biochemical methods for miRNA target identification [8]. It was the first experimental approach applicable to determine experimentally miRNA target sites in large scale. According to CLIP, in vivo protein-RNA complexes, here RISC-target interactions, are covalently crosslinked by 254 nm ultraviolet (UV) irradiation and then immunoprecipitated with an antibody to the Ago protein. Structure analyses have revealed that Ago makes sufficiently close contact to both miRNA and mRNA allowing for simultaneous identification of RISC-bound miRNAs and the target region on the mRNA, see figure 3.6(A). RNA not involved in the physical interaction with RISC is digest by ribonucleases. Crosslinked RNA including the miRNA and the mRNA fragment is sequenced with high-throughput methods. mRNA sequence reads are mapped to the genome to determine the mRNA identity as well as the site of Ago-mRNA interaction referred to as *Ago footprint*. Lastly, to determine which miRNA was bound to which target the 45 to 62 nt long Ago footprints are scanned for sites matching to seeds of crosslinked miRNAs. Hence, HITS-CLIP significantly diminishes the number of false positive predictions by providing Ago-bound miRNAs and Ago footprints. A drawback of this method is the low efficiency of UV 254 nm protein-RNA crosslinking. Increasing UV inten-

sity might improve sensitivity but on the other hand can result in DNA damage [115].

Hafner et al. presented PAR-CLIP (photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation) which is an improved method for isolation of RNA segments bound by proteins [9]. Cells are fed with photoreactive 4-thiouridine (4SU), which becomes incorporated into newly transcribed RNA in vivo with no apparent impact on transcript expression. Notably, 4SU substituted RNA is 100-1,000 fold more efficiently crosslinked than unsubstituted RNA. Further, crosslinking occurs already at longer wavelength (365 nm), i.e. cells are less prone to UV damage. In the preparation for sequencing the protein-bound RNA fragments are reverse transcribed to *complementary* DNA (cDNA). The reverse transcriptase incorporates G opposite to crosslinked 4SU instead of A. Consequently, the RNA sequence reads obtained from high-throughput sequencing showed a T to C mutation compared to the genome. Hafner et al. could show that T to C transitions can be used to map accurately the site of crosslinking between RNA and protein. 41 nt long regions centered over the predominant T to C mutation were extracted from clustered sequence reads that corresponded to RISC-bound RNA fragments, see figure 3.6(B). The length of the *crosslink-centered regions* (CCR) was chosen to include all possible mRNA-miRNA duplexes covering the T to C substitution respectively the crosslinking site.

3.2 Introduction

The most reliable indication of a miRNA-mRNA interaction is a perfect match between the mRNA sequence and the miRNA seed region of maximum length 8 nt. The corresponding part of the mRNA is referred to as the seed match, see figure 3.1. As introduced in section 3.1.2, different types of seeds have been determined based on evolutionary evidence [7]. Conserved seed matches associated to these classical seed types occurred more often in 3'UTR sequences than expected by chance. We have defined a minimal and sufficient set of seed types [1], see figure 3.8, based on experimentally identified miRNA target sites by Chi et al. using the HITS-CLIP method [8]. Our work complemented the work of the Bartel group as our division in functional and non-functional seed matches did not depend on the conservation of a match but was due to if a seed match was located within a Ago footprint or not. We confirmed that the seed region is confined to the first 8 nt at the miRNA 5' end and that seeds have a minimum length of 6 nt. Importantly, our study revealed the importance of the 6mer seed starting at position 1 of the miRNA. This seed type is not covered by the classical seed types.

Chi et al. did not provide the precise location of the Ago footprints but only the position of seed matches residing within Ago footprints. Therefore, the Ago footprint regions had to be estimated in order to define the surrounding non-footprint regions that harbor non-functional seed matches. Hafner et al. provided the precise chromosomal location of target sites distributed on the transcriptome of HEK293 cells [9]. This data set allowed for a more accurate definition of functional and non-functional sites.

The following project consists mainly of two parts. First, the seed matches were grouped to seed types ensuring maximum sensitivity, that is, all functional sites had to be taken into account. This is in contrast to the object of the Bartel group who required seed types to have high signal-to-noise ratios (i.e. specificity) [7]. Importantly, seed matches of the 3'UTR and the coding region were considered separately. Second, the resulting seed types were characterized and compared with each other in terms of several quantitative and qualitative features typically discussed in conjunction with miRNA target recognition: Each of the obtained seed types was examined for its sensitivity and specificity by means of the numbers of functional and non-functional sites associated to it. Further, seed type-specific effects on mRNA stability as well as preferences of miRNAs for certain seed types were explored. Finally, the conservation of seed matches was analyzed. In terms of the 3'UTR, the following is in some parts a recapitulation of the work published by Ellwanger et al. [1] but with a refined data basis. With regard to the ORF a comparable study has been not performed yet.

3.3 Methods and Materials

3.3.1 Preparation of miRNA-mRNA interaction maps

The miRNA-mRNA interaction data for the following analyses was obtained from Hafner et al. [9]. Using PAR-CLIP, Hafner et al. determined 17,318 CCRs. Each of which is supported by at least 5 sequence reads and 20% T to C transitions.

The mRNA data used was from the NCBI RNA reference sequences collection (RefSeq) [116] and were downloaded from UCSC [117] on September 2012. As Hafner et al. referred to genome assembly hg18, all sequence data was obtained from the hg18 database on UCSC. In case several annotated RefSeq sequences were available for a gene the respective longest transcript was selected. Further, only RNAs having a defined 5'UTR and 3'UTR and a coding region were considered. The final set consisted of 18,613 mRNAs that were screened for CCRs in the next step. CCRs were required to be completely contained in mature transcripts, i.e. within exonic regions. Further, CCR sequences that did not match perfectly to the corresponding part of the mRNA sequence were removed. In this way 13,695 CCRs could be mapped to 5,569 mRNAs. The CCRs were distributed on mRNA regions as follows, 5'UTR: 290 (on 266 RefSeq sequences), coding region: 6,870 (3,672) and 3'UTR: 6,535 (3,483).

Hafner et al. prepared profiles of endogenously expressed miRNAs by means of three different measurements. Based on total RNA isolated from HEK293 cells and miRNAs isolated from non-crosslinked AGO1-4 immunoprecipitations (AGO-IP), they determined profiles of non-crosslinked miRNAs. The profiles obtained by these two approaches showed a high correlation. The expression of crosslinked miRNAs was derived from the miRNAs present in the combined AGO1-4 PAR-CLIP library. Comparison of crosslinked miRNA profiles with the combined non-crosslinked libraries showed a good correlation. However, it is worth noting, that a group of miRNAs exhibited systematically lower expression in the AGO-IP samples than expected based on the AGO-PAR-CLIP data. In addition, the frequency of T to C mutations, which is an indication for crosslinking, was substantially lower for this subset of miRNAs. It was suggested to rely on the profiles from AGO-IP, since the miRNA profiles based on AGO-PAR-CLIP are to some extent distorted by the abundant miRNA background (Markus Hafner, personal communication, October 2012).

640 miRNAs are listed in the supplementary material of [9]. The miRNA sequence read counts of the AGO-IP 1-4 experiments were combined. 358 miRNAs had at least one sequence read in the combined AGO-IP. miRNA mature sequences were downloaded from miRBase (version 15) [118]. miRNAs with equal seed se-

quence (position one to eight in the seed region) were combined to seed families. Each miRNA seed family was represented by the miRNA exhibiting the highest AGO-IP read count. In total, 294 different miRNA seeds formed the initial set. Friedmann et al. [87] determined the degree of conservation of miRNA families. They classified miRNAs as "broadly", "intermediate" or "mammalian" conserved. 146 miRNAs could be assigned to one of these categories. The set was further reduced on the 58 most abundant miRNAs accounting for 95% of the total of combined AGO-IP miRNA sequence reads, see table 3.1. 51 of these miRNAs were conserved at least among mammalians.

To relate the miRNAs to the CCR-containing transcripts, the mRNA sequences were scanned for sites that perfectly match to the seed region with a minimum length of 6. Importantly, the retrieved seed matches were disjunct, i.e. only the longest match of a miRNA at a particular site was recorded. It was distinguished between seed match start and target site start. A seed match on the transcript was allowed to begin either at position 1, 2 or 3 relative to the 5' end of the miRNA. The corresponding target site start referred in any case to position one. Analyzing the positional distribution of target sites starts of the 58 highly expressed miRNAs within the 41 nt long CCRs revealed an accumulation of starts upstream of the predominant T to C position, see figure 3.7. The remaining lowly expressed miRNAs caused a less clear signal at this position. This finding agreed with [9] who described varying T to C mutation frequency around seed matches with one peak almost directly at the seed match end. The positional bias could be observed for both coding region and 3'UTR, although the signal was stronger in the 3'UTR, see figure 3.7(A). In particular, at position +8 a significant enrichment of target site starts could be observed in both mRNA regions (standard scores: 4.7 (3'UTR), 3.96 (coding region)). This observation suggested grouping of target sites into two categories: functional and non-functional sites. Sites starting at position +8 or at adjacent positions of +8 that showed a frequency above the mean in both 3'UTR and coding region were classified functional. Thus, an extended contiguous peak region was defined ranging from positions +6 to +10 for both 3'UTR and coding region. Target sites starting in the twilight zone, that is, within CCRs but outside of the peak region were ignored. All other sites were classified non-functional.

To be associated either with a 3'UTR or a coding region, the center of the CCR had to be located in the respective region. The 58 miRNAs were tested for enrichment in 3'UTRs or coding regions by comparing the numbers of functional and non-functional sites of each miRNA with the summed lengths of peak regions and not-CCR regions over all 3'UTRs or coding regions, respectively. The latter two numbers represent the background. Considering 3'UTR and coding region separately allowed checking if groups of miRNAs do possibly exist that have an ex-

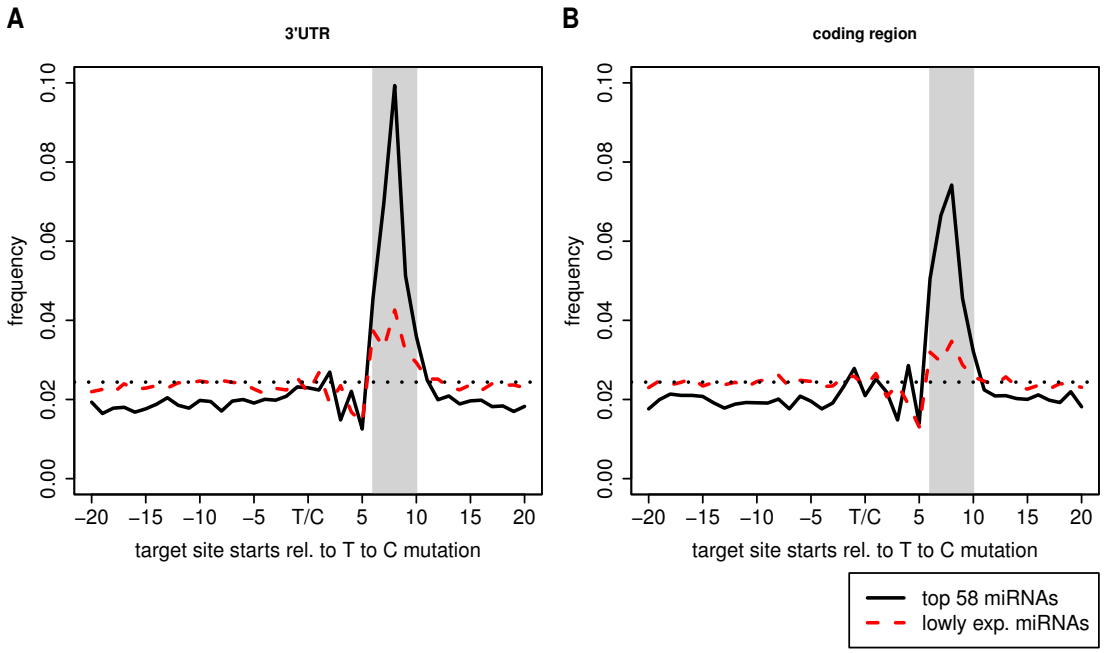


Figure 3.7: Positional frequencies of target site starts within CCRs. The dotted black lines represent the average frequency of target sites starts per position in CCRs located in the 3'UTR (A) and coding region (B), respectively. T/C indicates the predominant T to C transition within clusters of sequence reads and represents the center of CCRs. Target sites starting in the grey-shaded region were classified functional.

clusive bias to one of the two regions. It turned out that enrichment in the 3'UTR is highly correlated with enrichment in the coding region, PCC: 0.93, considering all 58 miRNAs. Applying two-tailed Fisher's exact test and multiple testing correction revealed 38 miRNAs to be significantly enriched in both regions, four miRNAs were biased to the 3'UTR and one miRNA was only enriched in the coding region. 15 miRNAs were not significantly over-represented in any region, see table 3.1. Concluding, based on this miRNAs do not seem to have different biases to mRNA regions. The set of miRNAs was further reduced on the 38 miRNAs that were significantly enriched in both mRNA regions.

Finally, a miRNA-ORF and a miRNA-3'UTR interaction map were obtained. The former included 1,580 coding regions, the latter contained 1,873 3'UTRs. Both contained 38 miRNAs. A mRNA region needed to have at least one functional site to be taken into account. Further, target sites starting too close (distance < 20) to the start codon (in case of the coding region) or to the stop codon (in case of the

	miRNAs
included in suppl. material of [9]	640
included in combined AGO-IP	358
unique seed region (nt 1 - 8)	294
95% of miRNA sequence reads	58
not enriched in any region	15
enriched in 3'UTR	4
enriched in ORF	1
enriched in 3'UTR and ORF	38

Table 3.1: miRNA sets. The first four rows illustrate how the 58 top expressed miRNAs were selected from the complete list of miRNAs provided by [9]. The 58 top expressed miRNAs were tested on over-representation in 3'UTR, coding region and both regions simultaneously.

3'UTR) were not considered in the respective set. These sites potentially contained seed matches that were not fully included in the respective mRNA region. The coding regions comprised 3,692 functional and 104,869 non-functional target sites. 4,880 functional and 129,625 non-functional sites were located on the 3'UTRs.

3.3.2 Conservation of sequence elements

The PhastCons method was used to determine the conservation of sequence elements. The underlying algorithm is based on a two-state ("conserved", "non-conserved") phylogenetic hidden Markov model that identifies evolutionary conserved elements in a multiple alignment [119]. Using a maximum likelihood approach, PhastCons estimates the models for conserved and non-conserved regions from the alignment. Next, the program scans along the alignment for regions that fit better to the conserved model than to the non-conserved model. Each nt of the reference sequence, in this case the human genome, gets a score representing the probability, that is a value between 0 and 1, to be part of a conserved element. For instance, Betel et al. have already used PhastCons scores to filter conserved miRNA target sites. If the score of a nt exceeded 0.57, the nt was considered to be conserved in mammals [120]. The conservation scores for the human genome (assembly hg18) based on a 17-way vertebrate alignment were downloaded from the UCSC database [117].

A sampling approach was used to test if a group of functional seed matches, e.g. sites based on the same seed type, was significantly conserved in a set of target se-

quences. Based on the target sequences a background set was compiled containing all sites of the same length as the considered seed match type. The conservation score of a functional or a background site was determined by the minimum PhastCons base score at that site. A subset of equal size as the set of functional sites was drawn without replacement from the background set. The Wilcoxon test was applied to test if the PhastCons score distributions of the sampled subset and the functional seed matches were significantly different from each other. Subset sampling and subsequent testing were repeated 10,000 times. The proportion of tests with P-value < 0.05 constituted the final P-value.

Further methods

Preparation and analysis of the data was done with the programming language Java and the data analysis workbench KNIME [121]. The R software was applied for statistical computations and plotting [122]. For hierarchical clustering the R package pvclust was used that allows for assessing the uncertainty of clusters with AU (Approximately Unbiased) P-values obtained via multiscale bootstrap resampling [123].

3.3.3 Evaluation measures

Seed matches were grouped by seed types and the resulting sets of sites were assessed by means of *sensitivity*, *specificity*, *precision* and the *Matthews correlation coefficient* (MCC). Functional sites included in a seed type-related set were considered as *true positives* (TP) and included non-functional sites were tagged *false positives* (FP). Further, *false negatives* (FN) were all functional sites and *true negatives* (TN) were all non-functional sites that were grouped to other seed types, respectively.

Statistical measures to assess the performance of binary classifiers have been used to characterize the "predictive power of seed types". The *sensitivity* of a seed type-related set indicates the proportion of functional sites covered by the set. *Specificity* denotes the proportion of excluded non-functional sites and *precision* represents the proportion of functional sites in a seed type-related set. Sensitivity, specificity and precision represent complementary properties of a classification system. The MCC indicates the correlation between the observed distribution of functional and non-functional sites, i.e. the background, and the distribution obtained for a certain seed type.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.4}$$

3.4 Results

3.4.1 Definition of miRNA seed types

A match between miRNA and target sequence was required to start either at position 1, 2 or 3, subsequently denoted α, β, and γ, relative to the 5' end of the miRNA. The search for perfect complementary sites to the 38 miRNAs with a minimum length of 6 nt resulted in a variety of match lengths. Matches of equal length sharing the same start position constituted a seed match type, see [1]. In total, 19 different seed match types were retrieved from the 1,580 coding sequences and 18 from the 1,873 3'UTRs, see table A3.5[1]. Adopting each of the 19 seed match types as a seed type would yield a very complex set of seed types compared to the 5 classical seed types proposed by the Bartel group [7]. Further, only 0.046% (3'UTR) and 0.033% (coding region) of the functional sites were based on seed matches of length 9 or longer. Hence, on the one hand seed match types with a length above 8 were only poorly represented, but on the other hand they made up almost half of the seed match types. It appeared reasonable to reduce this complexity by merging long but marginally represented seed match types with shorter ones that covered already a substantial portion of the seed matches.

But what is the optimal cutoff to differentiate between substantial and non-substantial seed match types? Ellwanger et al. proposed a procedure to address this problem. This method determines substantial seed types by means of their proportions of functional and non-functional sites, see section 3.1.4.

[1]The numbers of figures or tables included in the Appendix are preceded by an "A".

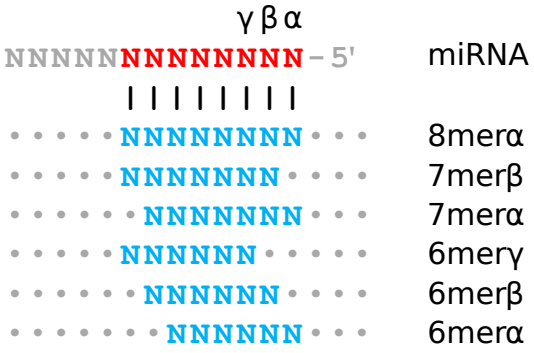


Figure 3.8: Minimal and sufficient set of seed types according to [1].

Seed type	Fun.	Non-fun.	LOR
6merα	24	32	-0.12
6merβ	19	21	-0.04
6merγ	24	28	-0.06
7merα	10	8	0.12
7merβ	13	8	0.21
8merα	9	3	0.44

Table 3.2: Sites per seed type. Proportion (%) of functional (Fun.) and non-functional sites in 3'UTRs and log adds ratio (LOR) according to [1].

3'UTR-related set of seed types

Figure 3.8 and table 3.2 introduce the seed type set proposed in [1]. Applying the above presented procedure for filtering substantial seed types on the miRNA-3'UTR interaction map yielded 11 seed types, see table 3.3. The stricter definition of functional sites applied in this work extended the set by the 7merγ and some seed types with length 8 and longer. On the other hand, functional 6mer sites, particularly, 6merα sites comprised a smaller fraction here.

Interestingly, the 8merα was stronger enriched among functional sites as indicated by the log odds ratio (LOR) than the longer types 9merα and 10merα. Commonly, one would expect that the proportion of false positives, i.e. non-functional sites, decreases with growing length of the sequence motif. A similar trend could be observed for seed types starting at the β-position. The enrichment of functional sites grew until 7merβ and subsequently decreased again. Both seed types starting at the γ position were under-represented with the shorter 6merγ type being more enriched than the 7merγ type among functional sites. Concluding, the decrease of precision, see table A3.7, for seed types going beyond position 8 of the miRNA sequence supports the conception of the seed to be formed within the first 8 nt. Further, the significant under-representation of 6mer seed types and seed type 7merγ points to positive selection of seeds with length ≥ 7 that start either at position 1 or 2 relative to 5' end of the miRNA.

The seed type set proposed by Bartel and colleagues [7] covered all functional seed matches except the 6merα sites, see figure A3.18. Bartel et al. required

	Functional		Non-functional			
Seed type	Freq.	%	Freq.	%	LOR	P-value
6merα	528	11	38,745	30	-0.53	$1.99E^{-204}$
6merβ	823	17	27,949	22	-0.13	$7.35E^{-015}$
6merγ	823	17	26,694	21	-0.10	$4.09E^{-010}$
7merα	577	12	11,319	9	0.14	$5.99E^{-012}$
7merβ	924	19	7,443	6	0.55	$1.02E^{-188}$
7merγ	243	5	9,741	8	-0.18	$1.48E^{-011}$
8merα	534	11	3,211	2	0.63	1.83^{E-145}
8merβ	212	4	2,085	2	0.42	$2.57E^{-031}$
9merα	125	3	1,008	1	0.49	$5.92E^{-025}$
9merβ	60	1	903	1	0.24	$1.52E^{-004}$
10merα	31	1	527	0	0.19	$2.43E^{-002}$

Table 3.3: SEED TYPES IN THE 3'UTR. The table shows the numbers of functional and non-functional seed matches grouped by seed types that have been identified by the method introduced in [1]. LORs and P-values are based on comparison with the background frequencies.

perfect complementarity of the target site to miRNA positions 2-7 or 2-8 but not to position 1. Therefore, the mapping between the two sets was not one-to-one. Each of the α and β seed types, except 6merα, overlapped with two of the classical seed types. For instance, 6mer and 7mer-A1 matches were grouped to seed types 6merβ and 7merα with 6merβ matches covering the larger proportion of 6mer sites and 7merα sites comprising the larger proportion of 7mer-A1 sites. 7mer-A1 matches are flanked by an adenine on the target that is opposed to miRNA position 1. As most of the miRNA sequences start with an uracil [5] (here: 24 of 38) the majority of 7mer-A1 seed matches equated 7merα sites and vice versa. In the event the miRNA starts not with an uracil, 7mer-A1 is equal to 6merβ. Consistently, the majority of 6merβ sites was based on 6mer matches that are not flanked by adenines on the target by definition. The partitioning of 7mer-m8 and 8mer sites can be explained analogously.

Figure 3.10(A) illustrates sensitivity and specificity of seed types. Although under-represented in CCRs, about 45% of functional sites were based on 6mers indicating the great importance of short seed types for sensitivity. Seed type 6merα played an exceptional role in this context: It was least frequently co-located with other seed types within 3'UTR sequences, see table A3.6. Importantly, the mRNAs

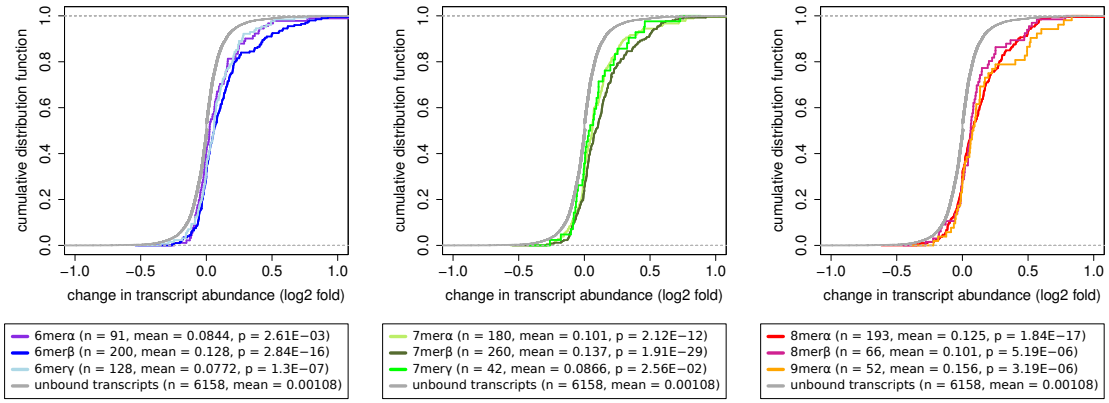


Figure 3.9: Effect of seed matches on transcript stability through the 3'UTR. Transcripts were categorized according to the seed types present in the 3'UTR. The transcript sets are not disjunct. Only transcripts having functional sites exclusively in the 3'UTR were used. Further, each 3'UTR needed to have at least one functional seed match to one of the 14 blocked miRNAs, see description in the text. The *P*-values indicate the significance of the difference between the changes of transcript levels of transcripts containing CCRs versus transcripts without CCRs. *Left:* 6mer seed types, *Center:* 7mer seed types, *Right:* 8mer and longer seed types.

used in this analysis included functional sites exclusively in the 3'UTR. 37% of the 3'UTRs containing a functional 6merα site did not exhibit functional seed matches associated to other seed types. For all other seed types the fraction of exclusive targets was considerably lower. The same holds when focusing on CCRs: Here, less than half of the CCRs with a functional 6merα seed match contained other seed types. Consequently, seeking for 6merα sites is necessary as it holds many exclusive targets.

On the other hand short seed types were less suited to discriminate between functional and non-functional sites as they involved many false positives. The combined specificity of short seeds amounted to 0.2. Each of them was below the dashed line in figure 3.10(A) evincing a performance even worse than random guessing, that is, inverting the classification obtained by 6mer seeds would have yield a result superior to average random prediction. The remaining proportion of functional target sites included seeds of length 7 or longer. Two of them 7merβ and 8merα stood out significantly when considering the MCC, see table A3.7. 7merβ represented the best classifier closely followed by 8merα. The combined set of 7merβ and 8merα seed matches achieved a sensitivity of 0.3 and a specificity

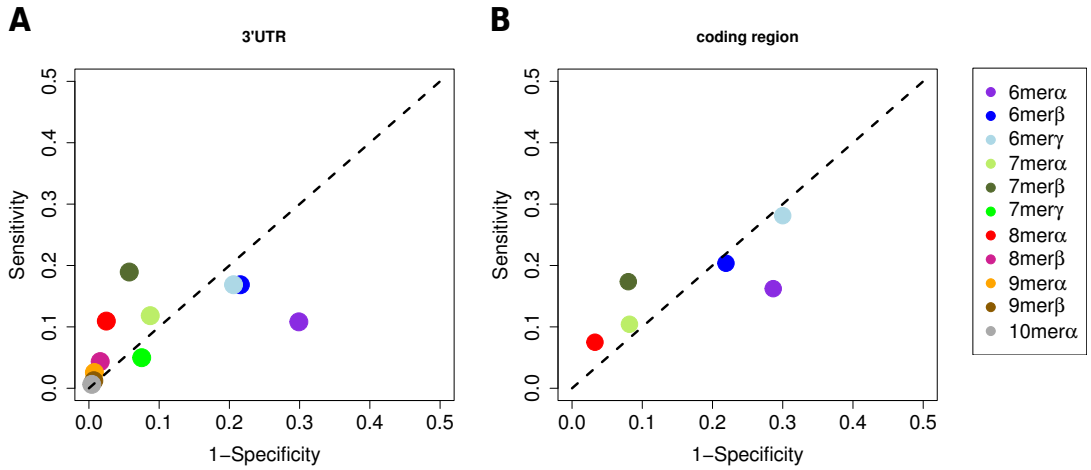


Figure 3.10: PREDICTIVE POWER OF SEED TYPES. The performance of each seed type as classifier was evaluated by means of sensitivity and specificity based on seed matches from 3'UTR (A) and coding region (B). The dashed black lines represent average random prediction, respectively.

of 0.82. 7merα contributed appreciable to sensitivity but was less specific. 7merγ was clearly lagging behind these three. It exhibited the third worst prediction performance of all seed types according to the MCC.

Beside the sequence signal properties, the repressive effect of seed types on mRNA expression was quantified. Figure 3.9 displays the change of transcript abundance in dependence on the seed matches occurring in the respective 3'UTRs. The expression data was taken from Hafner et al. [9]. They blocked the most abundant miRNAs in HEK293 cells and compared the change of transcript abundance with mock transfected cells. 14 of the 38 miRNAs considered here belonged to the group of blocked miRNAs. The set of 3'UTRs was filtered for this analysis: Only transcripts were taken into account that had CCRs exclusively in the 3'UTR. Further, only functional matches of the blocked miRNAs were regarded. Consequently, each of the 3'UTRs used for this analysis had at least one functional seed match to one of the 14 blocked miRNAs. The transcripts were grouped to subsets according to the seed types present in their 3'UTRs. For instance, the set of 6merα included all transcripts that had minimum one functional 6merα seed match. The transcript sets were not disjunct. Non-overlapping sets would have been too small for the analysis as the majority of 3'UTRs contained matches of different seed types, see table A3.6. Hence, for most transcripts the obtained repression was not only caused by the denoted seed type. In addition, further factors such the number of target sites can affect the extent of expression, as well [94]. But this applied

	Functional		Non-functional			
Seed type	Freq.	%	Freq.	%	LOR	P-value
6merα	599	16	30,039	29	-0.31	1.40E^{-063}
6merβ	752	20	22,983	22	-0.04	3.02E^{-002}
6merγ	1,038	28	31,454	30	-0.04	1.83E^{-002}
7merα	385	10	8,573	8	0.11	5.36E^{-006}
7merβ	641	17	8,399	8	0.36	2.06E^{-066}
8merα	277	8	3,421	3	0.36	2.74E^{-031}

Table 3.4: SEED TYPES IN THE CODING REGION. The table shows the numbers of functional and non-functional seed matches grouped by seed types that have been identified by the method introduced in [1]. LORs and P-values are based on comparison with the background frequencies.

to each of the target sets and therefore an impression of the repression strength could be obtained.

Each set was compared to the change of expression of transcripts containing no CCRs, denoted "unbound transcripts". Among the 6mer seeds, 6merβ reduced mRNA stability strongest followed by 6merγ. Transcripts containing 6merα sites were only weakly but still significantly repressed (Wilcoxon test with Bonferroni correction, P-value < 0.05). Importantly, functional 6merα matches were least frequently co-located with sites of other seed types, see table A3.6, i.e. 6merα-based transcript destabilization gained least support by other seed types. Since 10merα and 9merβ sites were rare they have been merged with 9merα and 8merβ, respectively. Concluding, transcripts including 9merα, 7merβ, 6merβ or 8merα sites were on average most affected.

ORF-related set of seed types

Generally, the LOR values of the seed types obtained for the coding region were closer to 0, see table 3.4. Functional 6merβ and 6merγ matches were almost as frequent as it could be expected from the background distribution, but they were still significantly under-represented. The selective pressure to preserve the encoded amino acid sequence affects the evolution of miRNA target sites in the coding region [124]. Indeed, most amino acids are encoded by several codons leaving room for sequence variation. But as demonstrated in section 3.4.3 the conservation of 3'UTRs is much lower and thus allows for faster adaptation to miRNA targeting.

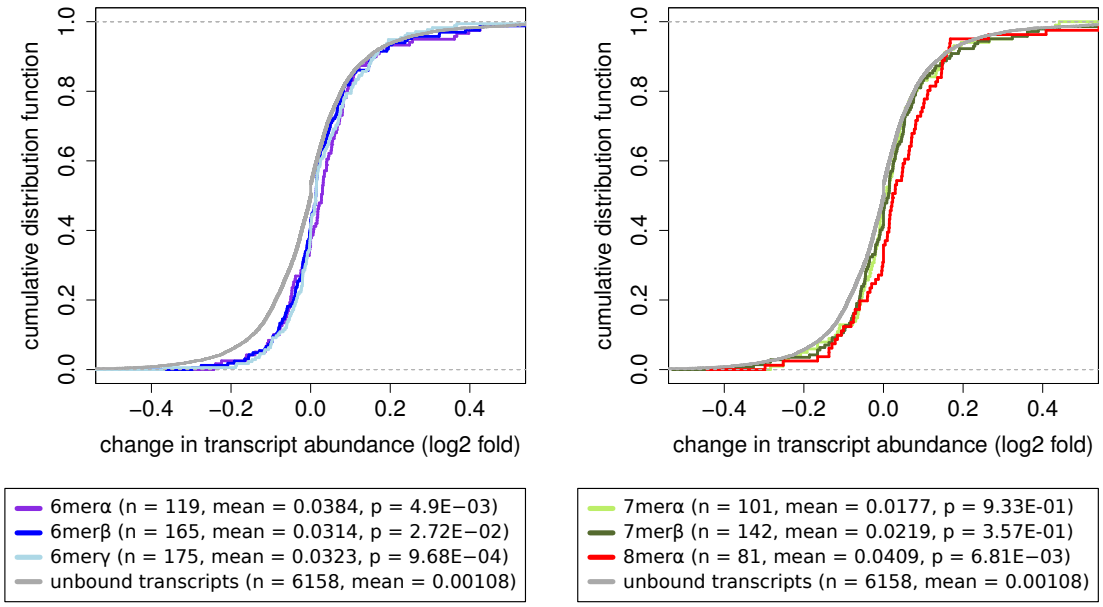


Figure 3.11: Effect of seed matches on transcript stability through the coding region. Analogous to figure 3.9. Note: the ranges of the x-axes are different to figure 3.9 *Left:* 6mer seed types, *Right:* long seed types.

In figure A3.19 the proportions of functional sites per seed type is compared between coding region and 3'UTR. Seed matches of seed types that were exclusive for the 3'UTR have been merged depending on the start position with the respective longest seed type present also in the ORF-seed type set, i.e. with 6merγ, 7merβ or 8merα. The distributions varied significantly from each other (χ^2 test, P-value = 9.64E^{-048}). Surprisingly, a significant increase of short seed matches could be observed. Almost two-thirds of functional sites were of length 6 opposed to 45% in the 3'UTR or 50% when 7merγ sites were incorporated into 6merγ, respectively. Consequently, in the ORF short seeds had an even greater importance for sensitivity than in the 3'UTR, see also figure 3.10(B). Similar to the 3'UTR, the 6merα exhibited a strong preference to occur solely in coding regions, see table A3.6. On the other, hand short seed types were associated with many false positive matches. The proportions were similar as in the 3'UTR and therefore led to similar specificity values (specificity of combined short seed types: 0.19). While the MCC values increased for the short seeds, they decreased for the long ones, see table table A3.8. But nevertheless, 7merβ and 8merα again achieved the highest MCC values with 7merβ performing better than 8merα. The combination of these two reached a sensitivity of 0.25 and a specificity of 0.89.

miRNA-mediated regulation through the coding region has been described to be less effective as through the 3'UTR, see section 3.1.5. The impact of ORF-located functional sites categorized by seed types on transcript stability is shown in figure 3.11. Here, the used transcripts had exclusively functional seed matches in the coding region. Surprisingly, the change of expression of transcripts containing 7merα or 7merβ was not significantly different from that of unbound transcripts. This was noteworthy in particular with regard to 7merβ that has been one of the most effective seed types in the 3'UTR. Provided the diminished repressive effect through 7mers in the coding region, it was surprising that all of the short seeds reduced the expression level of their targets significantly with 6merα having been the most effective one, see mean values in figure 3.11.

Fang et al. [103] and Schnall-Levin et al. [104] found evidences for coordinated regulation between 3'UTR and coding region. They measured an increased repressive effect if both 3'UTR and coding region of a transcript contained target sites. Therefore, target sites in the coding region were characterized to enhance the effect of sites in the 3'UTR. Importantly, here the majority, 52%, of transcripts with functional sites in the ORF had no CCRs in the 3'UTR. Therefore, most of the detected ORF sites did not appear to support concurrent regulation through the 3'UTR.

3.4.2 Seed type preferences of miRNAs

An obvious question was now, after seed types were determined, if miRNAs have different preferences for seed types or if the individual miRNA seed type distributions were consistent with the basic populations as shown in tables 3.3 and 3.4 (see columns labeled "%" in the "Functional" sections) [125].

Only functional sites were used in this analysis. To facilitate the comparison between both mRNA regions, the long 3'UTR-unique seed types were merged with 6merγ, 7merβ and 8merα, respectively. First, each miRNA was tested if its seed type distribution matched the distribution of the basic population (χ^2 *Goodness of Fit* test and correction for multiple testing, P-value < 0.05). 31 of 38 miRNAs showed independent seed type distributions regarding the 3'UTR. 22 miRNAs differed with respect to the coding region. Hence, the basic distribution of seed types is generally not representative for individual miRNAs. The greater number of independent distributions with respect to the 3'UTR is possibly due to the lower selective pressure on 3'UTR sequences. As stated already, presumably, the 3'UTR was able to adapt more to miRNA-mediated regulation.

Clustering by means of the seed type frequencies revealed three clusters of miRNAs with respect to the 3'UTR, subsequently called A, B and C, and two groups

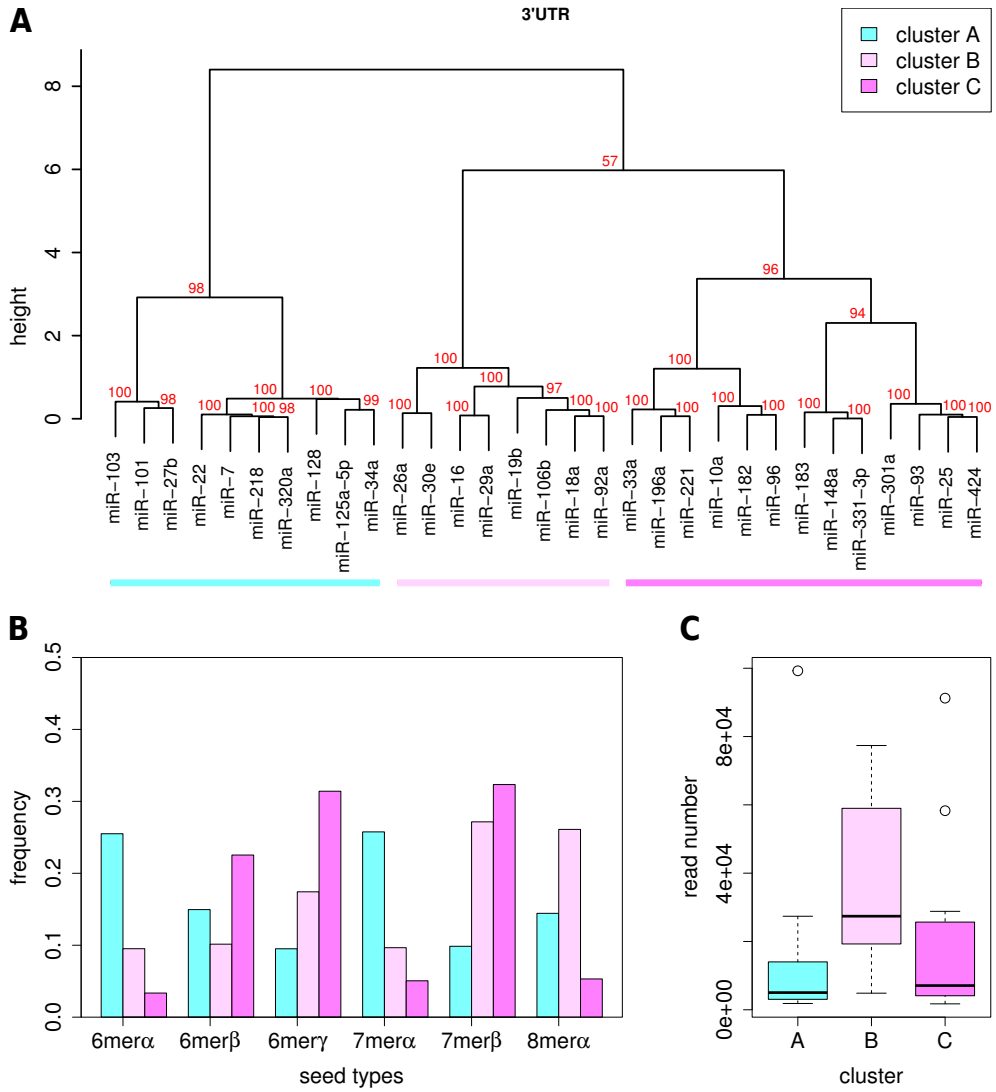


Figure 3.12: Seed type preferences of miRNAs (3'UTR). (A) miRNAs were clustered by means of their seed type frequencies regarding the 3'UTR (distance: PCC, clustering: Ward's method). The red numbers represent the probability values calculated by *pvclust* [123] indicating how strong the clustering is supported by the data. (B) Frequencies of seed types for the clusters shown in (A). (C) miRNA read numbers were categorized according to the clusters of (A).

of miRNAs related to the coding region, called D and E, see figures 3.12 and 3.13, respectively. miRNAs of cluster A were biased to seed types involving the α-position. 66% of the sites of cluster A belonged to α-seed types in contrast to

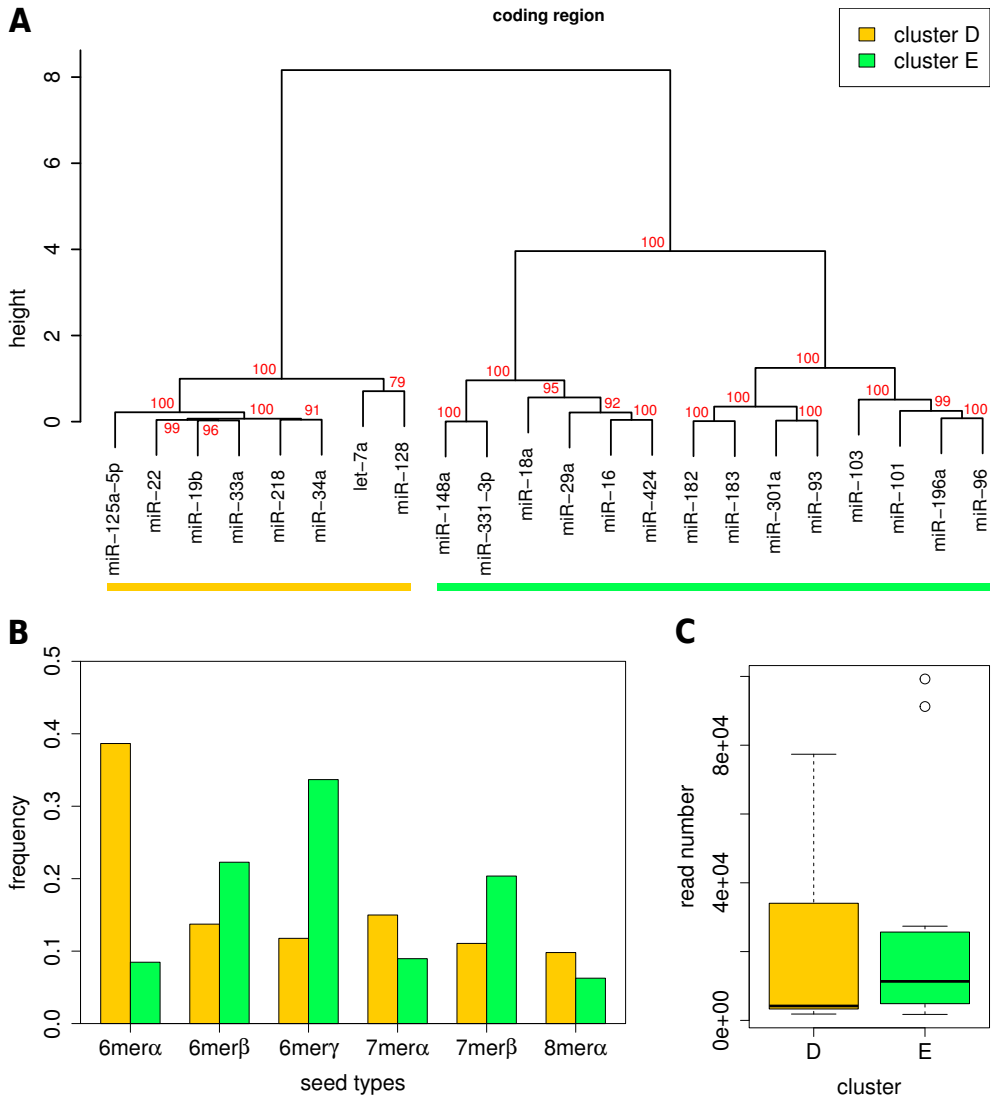


Figure 3.13: Seed type preferences of miRNAs with (coding region). Caption is analogous to figure 3.12.

only 13% regarding cluster C. On the other hand, 6merγ and 7merβ that were preferred by miRNAs of cluster C were under-represented in cluster A. Hence, miRNAs of clusters A and C exhibited very contrary preferences for seed types (PCC: -0.84). Regarding miRNA concentration, apart from some exceptions both groups were rather lowly expressed compared to cluster B. 53% of the functional sites of cluster B were either of type 7merβ or 8merα. High cellular concentration in combination with the two most efficient seed types suggested that miRNAs of cluster B had most impact on gene expression.

Except let-7a, the set of miRNAs with independent seed type distribution with respect to the coding region was a subset of the miRNAs found for the 3'UTR. Here, the clustering uncovered two main groups, see figure 3.13. Cluster D overlapped predominately with cluster A. Similar to cluster A, cluster D had also a strong bias to α-seed types, but here this preference was mainly due to seed type 6merα that made up 39% of the sites of cluster D. 9 of 14 miRNAs included in cluster E could be found in cluster C. Consistently, cluster E showed also an avoidance of the α-position although not to the same extent as seen in C (24% of the sites involved the α-position). Hence, cluster B appeared to be not represented in the coding region: 4 of 8 miRNAs of cluster B had no independent seed type distribution with respect to the coding region. Further, 10 of the 31 3'UTR-independent miRNAs had a significantly different seed type distribution regarding the coding region. 5 of them belonged to cluster B. The disappearance of cluster B that was biased to 7merβ and 8merα fitted well with the observation that long seed matches were less frequent in the coding region, see figure A 3.19.

Concluding, it could be shown that miRNAs may have different seed matching preferences. Highly expressed miRNAs were biased to long seed matches in the 3'UTR that induce strong target repression, whereas seed matches of less abundant miRNAs belonged to seed types affecting the target level only mildly in general. Abundant miRNAs were less influential in the coding region. It would be interesting to check if seed type preferences of miRNAs are condition-specific.

3.4.3 Conservation of miRNA targets sites

Sequences that are conserved among species are in general more likely to be functional than non-conserved ones. It is well known that miRNA targets with conserved seed matches in the 3'UTR are stronger regulated than those with non-conserved seed matches, see e.g. [94]. In the following the conservation of seed matches was explored. To determine the conservation of sites, the conservation scores computed by the PhastCons algorithm based on a 17-way alignment of vertebrates were used. This strategy to assess conservation of seed match sites has been used already by Betel et al. [120]. The PhastCons method produces a base-by-base conservation score denoting the probability to be evolutionary conserved across 17 vertebrates. The data was downloaded from UCSC [117], http://genome.ucsc.edu/.

Conservation of the CCRs in the 3'UTR Figure 3.14 displays the sequence conservation in the immediate vicinity of CCRs located in 3'UTRs. Conservation scores of the positions 100 nt upstream and downstream of the crosslinking site,

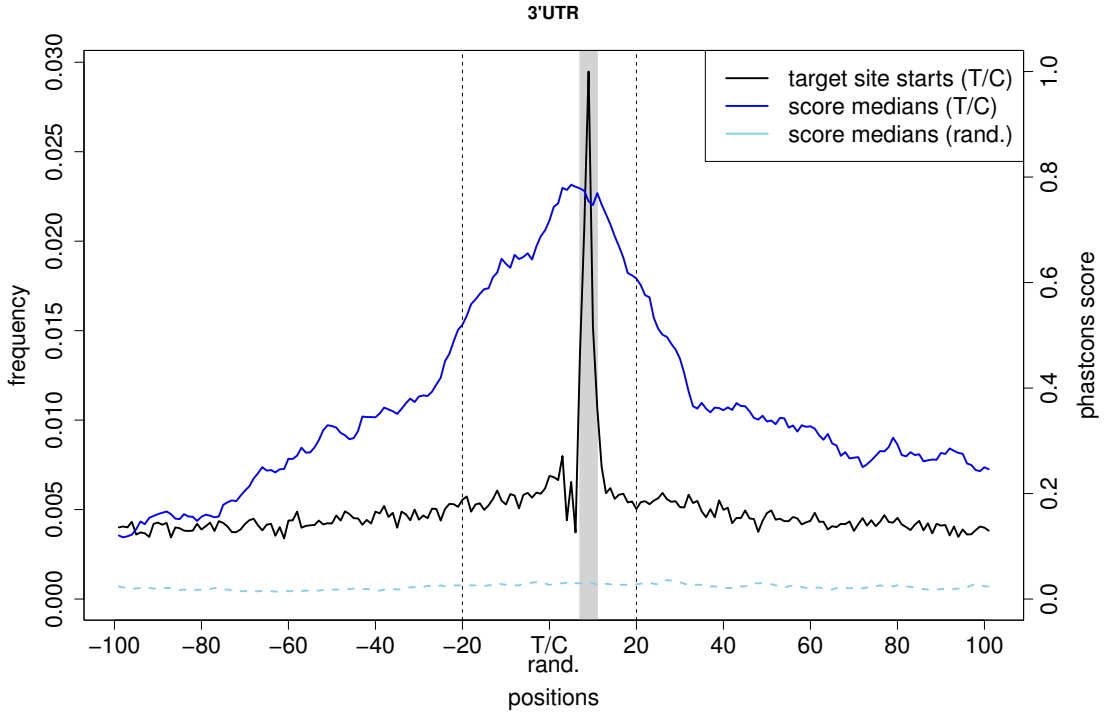


Figure 3.14: Conservation of CCRs in the 3'UTR. For 5,382 CCRs the conservation scores of the nt upstream and downstream of the T to C transition were collected, respectively. The medians of the PhastCons scores are shown by the dark blue line. Conservation scores surrounding 6,397 randomly selected positions in the 3'UTR are represented by the light blue line. Frequencies of target site starts per position of highly expressed miRNAs are indicated by the black line, compare with figure 3.7. Target sites starting in the grey-shaded region were classified functional. The dashed black lines show the borders of the CCR. Although not clearly visible here, similar to the extended CCRs, the scores based on randomly selected 3'UTR fragments were rising towards the end on the right (downstream of the T/C position).

i.e. the T/C position, were recorded. The darker blue line displays the medians of the resulting score distributions for each position. To get an impression of the basic conservation of 3'UTR sequences, the conservation around randomly selected positions is shown as well (light blue line). It was required that the 201 nt long regions both the extended CCRs and the randomly determined fragments used for this analysis had to be fully inside the 3'UTR. Therefore, not all of the 6,535

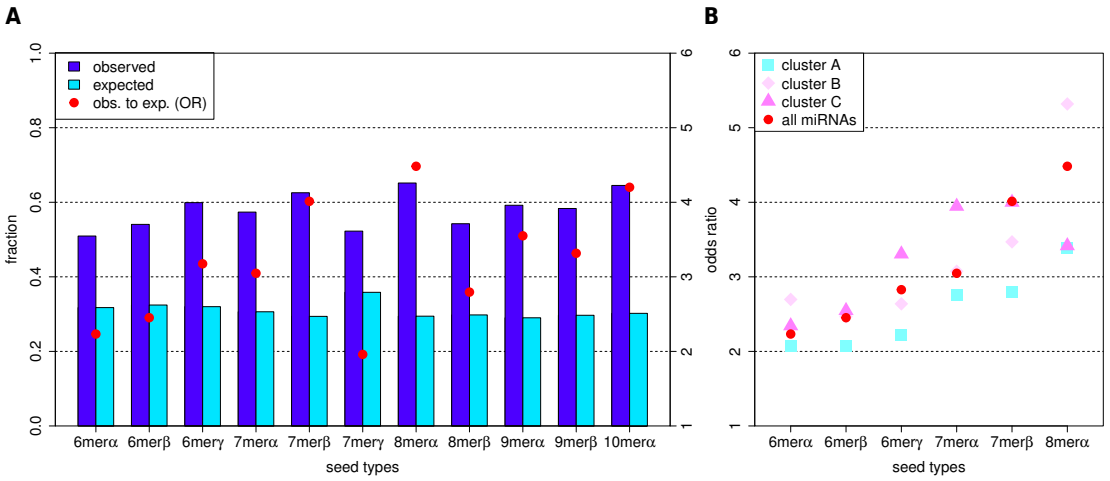


Figure 3.15: Conservation of functional sites in the 3'UTR. (A) Observed and expected fraction of conserved seed matches in the 3'UTR categorized by seed types. Red points denote the ORs. (B) ORs for the miRNA clusters, see section 3.4.2.

CCRs located in 3'UTRs could be considered here.

Three sections of distinct conservation could be identified. The seed match region was most conserved. It is located between the T/C position and the right end of the grey colored peak region. After a decline upstream of the seed match, the conservation remained constant until position -11 relative to the T/C position. Presumably, here pairing between the 3' end of the miRNA and the target site is reflected, i.e. the target site ranges approximately from position 8 to -11. The last striking section was between positions -28 and 33. From both points the conservation scores rapidly increased towards the center. Possibly, these positions represent the borders of the Ago footprint that has been defined by Chi et al. [8]. They found that the region of interaction between the RISC and the target sequence spans 62 nt on average.

Conservation of seed match sites in the 3'UTR For each seed match type t, see table A3.5, a specific background set b was compiled to determine the expected conservation: f represents the set of functional sites of t. First, all 3'UTRs were collected that contained an elements of f. Based on this set of sequences, the background set b was created by recording all sub-sequences of the same length as the seed match type t. The minimum conservation score assigned to a base of a functional seed match or to a sub-sequence of the background set was used

as representative score for the complete seed match or sub-sequence, respectively. The resulting score distribution based on the set of background sub-sequences was bimodal with the plethora of sites having score 0 and a second peak at 1.0, see figure (A) of A3.20. Two analyses were conducted: For each seed type it was tested if the related set of seed matches was stronger conserved than the respective background by means of the PhastCons score distributions. Further, the conservation of seed matches was compared between seed types.

Except 10merα, the conservation score distributions were significantly different (P-value < 0.05) from the 3'UTR background, see table A3.9, that is, functional sites were stronger conserved. The conservation of seed matches of different types was compared relative to a reference score. In this work 0.57 was used that roughly corresponds to conservation across mammals, see section 3.3.2. Seed matches as well as background sites with a score ≥ 0.57 were classified conserved and the remaining sites were considered non-conserved. Sites of seed match types beyond the longest seed types, i.e. 10merα, 9merβ and 7merγ, were merged with these according to the start position. Figure (A) of 3.15 shows the fractions of conserved functional sites (observed) and the particular expected fractions. The expected fractions varied between the seed types. To account for these variations, the odds ratios (OR) were additionally calculated, see equation 3.5 (*cons.* gives the proportion of conserved sites of a set). The OR values were interpreted as indicators of the extent of purifying selection.

$$OR = \frac{\frac{cons.(f)}{1-cons.(f)}}{\frac{cons.(b)}{1-cons.(b)}} \tag{3.5}$$

Among 6mer sites, 6merγ seed matches turned out to be conserved most above the expected fraction even more than the longer 7merα sites. The group of 7merγ sites exhibited the lowest OR of all seed types. 7merβ and 8merα targeting appeared to be most subjected to purifying selection. All seed types beyond 8merα were less conserved above the background than 8merα. Although 10merα achieved a high OR value it is of less importance due to the small number of functional seed matches, see table 3.4. Hence, the decline of conservation beyond 8merα supported the common seed model stating that the seed region is confined to the first 8 nt of the miRNA. Further, conserved targeting was strongly based on seed types 7merβ and 8merα while the short seeds 6merα and 6merβ were more involved in non-conserved regulation. Removing non-conserved sites improved the specificity of all seed types, see figure (A) of 3.16. In total the number of non-functional sites dropped by a factor of 3.3 while the number of functional sites decreased only by factor of 1.7. Especially, the specificity of the 6mer seed types benefited from this

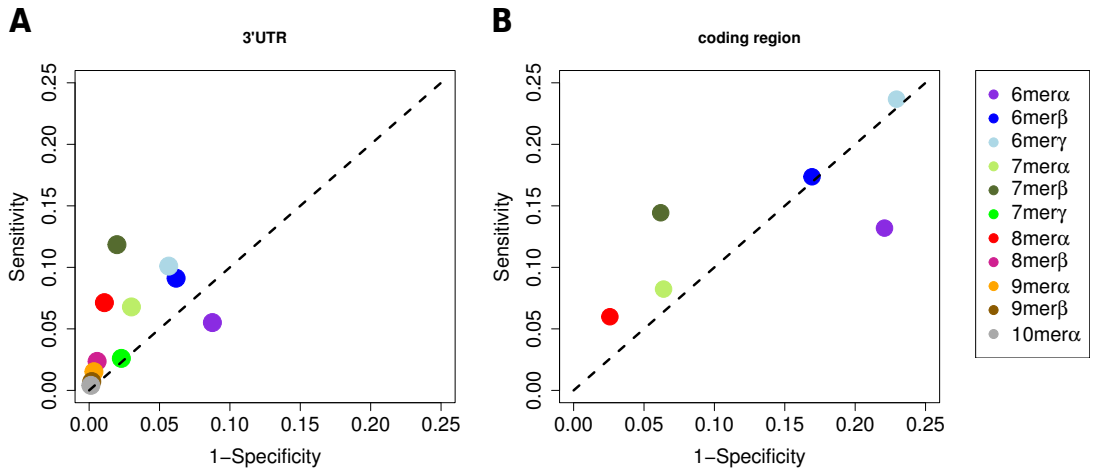


Figure 3.16: PREDICTIVE POWER AFTER REMOVING NON-CONSERVED SITES. The usage of seed types as classifiers was evaluated by means of sensitivity and specificity based on seed matches from 3'UTR (A) and coding region (B). Only sites with conservation score above 0.57 were retained. The dashed black lines represent average random prediction. Note: the ranges of the axes are different to figure 3.10.

filtering. Using only conserved 6merγ and 6merβ sites for target site prediction would cause a classification better than random guessing.

Figure 3.12 shows the miRNA clusters based on seed type preferences. Now, separately for each cluster the proportion of conserved seed matches per seed type was determined. Figure (B) of 3.15 indicates the obtained OR values. The seed matches of cluster A were least conserved above the background across all seed types. Nevertheless, the conservation score distributions were significantly higher than the background for each seed type, see table A3.10. Cluster A miRNAs had a preference to involve the α-position, but except 8merα, α-seed matches were not strikingly more conserved. The PCC was calculated to measure the relationship between seed type preference (represented by the numbers of conserved sites per seed type) and conservation (represented by the OR values). Seed type preference and conservation were not correlated for cluster A (PCC: 0.021). Opposed to cluster A, α-sites were under-represented for miRNA of cluster C. Consistent with this, the PhastCons score distribution of 6merα matches was not significant here. Further, the OR values suggested that 8merα matches were less subject to purifying selection than 7merβ matches. Therefore, a weak correlation could be found (PCC: 0.287). Cluster B was biased to seed types 7merβ and 8merα. A PCC of 0.803 indicated that seed matches of the preferred seed types were par-

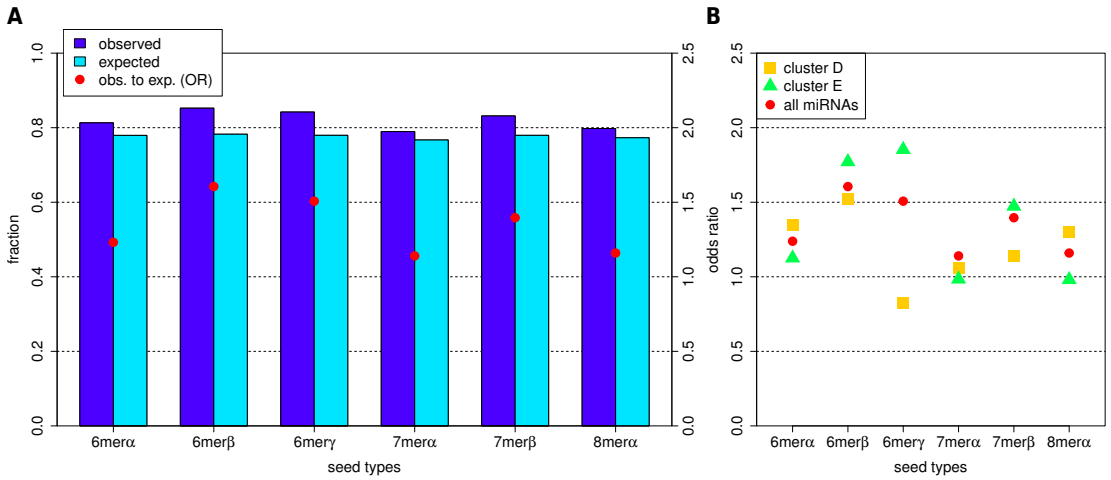


Figure 3.17: Conservation of functional sites in the coding region (A) Observed and expected fraction of conserved seed matches in the ORF categorized by seed types. Red points denote the ORs. (B) ORs for the miRNA clusters, see section 3.4.2. Note: the ranges of the axes indicating the ORs are different to figure 3.15.

ticularly conserved for this cluster. In summary, partially the different seed type preferences were reflected in the conservation of the seed matches. This applied in particular to the highly expressed miRNAs of cluster B that interact predominantly with conserved 7merβ and 8merα seed matches. On the other hand, the preferences of cluster C and A, both including mainly lowly expressed miRNAs, were only little or not at all conserved, respectively.

Conservation of seed matches in the coding region Due to the stronger selective pressure on the amino acid sequence, the coding region is much more conserved as the 3'UTR, see figure (B) of A3.20. Surprisingly, the functional sites of two long seed types, 7merα and 8merα, turned out to be not significantly conserved compared to the background, see table A3.9. On the other hand, the PhastCons score distributions of all short seeds differed significantly from the corresponding background distributions. Figure 3.17(A) shows the fractions of functional (observed) and expected sites conserved in mammals for each seed type and the obtained OR values. 6merβ and 6merγ matches achieved the highest OR values suggesting these matches are preserved most by purifying selection. Among the long seeds, 7merβ sites appeared to be most conserved. Remarkably, the proportion of 6merα seed matches exceeded that of both 7merα and 8merα sites. Already the frequency of functional sites per seed types pointed to an increased importance of shorts seeds

in the coding region, see figure A3.19. Both OR values and P-values as shown in table A3.9 implied that complementarity to the α-position is less necessary in the coding region. This applied in particular to matches of length > 6 as 6merα was both frequently occurring and significantly conserved in the ORF.

As expected, the effect of removing conserved sites was not as strong as in the 3'UTR due to the high basic conservation of the coding region, see figure (B) of 3.16. All in all, the number of functional sites decreased by a factor of 1.2 and the non-functional sites were reduced only by a factor of 1.3. As well as for the 3'UTR, selecting conserved 6merβ and 6merγ matches resulted in a classification superior to average random prediction.

Based on the coding region two miRNA clusters revealed, D and E. Seed type preferences of cluster D were not well recovered in the site conservation, see figure (B) of 3.17. The correlation between seed type preferences and conservation of seed matches was low (PCC: 0.302). Further, none of the conservation score distributions stood out from the respective background distribution, see table A3.10. Not so with cluster E, this had been biased particularly to seed type 6merγ. The conservation scores of 6merγ sites significantly differed from the ORF background. The same applied to 6merβ sites that were enriched as well in cluster E. On the other hand, seed types including the α-position were under-represented in cluster E. In agreement with this, the scores were not significant and the OR values were comparably small. In summary, again the seed types preferences of the cluster containing more abundant miRNAs, E, were stronger reflected in the conservation. Unlike the 3'UTR, here the tendencies to 6merγ and 6merβ sites emerged to be the strongest conserved preferences.

3.5 Conclusion

In this chapter rules and properties of miRNA seed-based target recognition have been explored based on experimentally identified miRNA target sites. Unlike the studies of the Bartel group [7], the data set used here was not biased to conserved regions and thus allowed for a more sensitive analysis. Further, conservation could be studied as a feature of seed matches. With respect to the 3'UTR, the higher resolution in the determination of functional seed matches resulted in a more complex set of seed types compared to the preliminary study by Ellwanger et al. [1]. Here, seed types exceeding position 8 relative to the 5' end of the miRNA showed significant individual proportions of functional and non-functional sites. However, the two longest seed types that ended exactly at position 8, 7merβ and 8merα, exhibited the best trade-off between sensitivity and specificity. Therefore, consistent with previous work but based on a different data basis this analysis

presented support for the notion that the seed region is confined to the first 8 positions of the miRNA.

The study has shown that focusing on the most specific seed types, 7merβ and 8merα, would imply to ignore a substantial fraction of functional seed matches. 50% of the functional sites were based on short 6mer seed matches in the 3'UTR. But 6mer matches were also found numerously outside of CCRs and they were less frequent conserved than long matches (except 6merγ that had more conserved sites than 7merα). Further, it was found that miRNAs adopt seed types differently. The preference for seed types depended on the abundance of miRNAs in the cell. Roughly summarized, highly expressed miRNAs were biased to long matches while less expressed ones tended to short seed matches. All in all, abundant miRNAs preferred long seed types that in turn were most frequently conserved and that showed the strongest effect on target stability. Importantly, short functional sites were stronger conserved as the background as well and had also significant impact on target expression. But with regard to the 3'UTR they performed worse in all respects, i.e. signal specificity, conservation, mRNA destabilization and expression of associated miRNAs, than the long seed types. Therefore, they have been less regarded by the miRNA target prediction community so far. In general, developers of target prediction methods aim to generate very specific predictions that have a high chance to be confirmed in subsequent experiments.

But the large number of short seed matches revealed here argues not to ignore them. Possibly, their primary function is not mRNA down-regulation. Possibly, transcripts containing short seed matches act as regulators of miRNAs, that is they function as ceRNAs. By interacting with miRNAs, ceRNAs keep miRNAs from regulating other mRNAs, see section 3.1.6. Against this background, the described properties of 6mer seed matches could be coherently explained. Unlike the specific miRNA-target interaction that should result in an efficient suppression of target expression, it is likely not that decisive for successful ceRNA-mediated control which mRNA molecule is acting as ceRNA. The direction of the miRNA-mRNA interaction is reversed, that is, the miRNA becomes the target of many ceRNAs. Quite in line with this, 6mer matches are frequent and additionally they are sufficient to form experimentally demonstrable interactions with mRNAs. As short seed matches occur numerously simply by chance there is less necessity to preserve them as strong as specific miRNA-target interactions by purifying selection. Further, the expression of ceRNAs is only marginally affected by 6mer matches, whereas the "authentic targets" [110] of ceRNA-sponged miRNAs are defended from degradation. Interestingly, seed type 6merα occurred least frequently with other seed types. According to this hypothesis, mRNAs containing exclusively 6merα sites would act only as ceRNAs. While other mRNAs that contain both

short and long seed matches were dependent on the present miRNAs either targets or ceRNAs or both. For instance, if certain circumstances induced a change of the protein expression pattern in the cell, the pool of interaction partners would change for miRNAs whose activity needed to be switched off or limited. (m)RNAs would increasingly occur that carry out ceRNA function with respect to these miRNAs. Conducting at this state an analysis such as in section 3.4.2 would reveal a preference for shorts seed matches of ceRNA-bound miRNAs. In parallel to regulation through ceRNAs, the miRNAs to be limited in their function are potentially subject to degradation and thus less abundant in the cell.

The investigation of miRNA-mediated regulation through the coding region is gaining importance, see section 3.1.5. Target sites in the coding region were reported to enhance parallel targeting through the 3'UTR. But importantly, this analysis showed that most of the transcripts containing target sites in the ORF were free of functional sites in the 3'UTR. Therefore, miRNA-mediated regulation through ORF-sites appears to be independent of the 3'UTR, at least in most cases. Analogously to the 3'UTR, seed types were defined for the ORF. Interestingly, a significant bias to short seed matches was found in the coding region. The proportion of functional sites was higher in the ORF than in the 3'UTR for each of the three 6mer seed types. Further, for all types of functional 6mer sites the conservation differed significantly from the background. Beside that, merely 7merβ sites showed significant conservation. Consequently, 7merα and 8merα seem to have only little importance in the ORF. Consistently, Marin et al. reported about reduced specificity of 8merα sites in the ORF [126]. But they did not analyze all kinds of 6mer seed types considered here.

Seed type preferences of miRNAs were less clear in the coding region. Abundant miRNAs tended primarily to 6merγ followed by 6merβ and 7merβ. Less frequent miRNAs were biased to 6merα sites. As generally the repression through ORF sites is weaker it comes to the question whether and if so how it could be distinguished between authentic targets and ceRNAs here? Do 6merα sites, that also in the ORF were least frequently co-located with other seed types, occur primarily on ceRNAs and 6merγ, 6merβ and 7merβ sites on authentic targets?

3.6 Appendix

		3'UTR			coding region		
Start	Len.	Fun.	Non-fun.	P-value	Fun.	Non-fun.	P-value
α	6	528	38,745	$8.81E^{-275}$	599	30,039	$8.43E^{-069}$
	7	577	11,319	$6.90E^{-070}$	385	8,573	$1.15E^{-012}$
	8	534	3,211	$1.51E^{-007}$	212	2,538	$4.31E^{-001}$
	9	125	1,008	$1.67E^{-004}$	47	651	-
	10	23	429	$3.45E^{-001}$	11	176	-
	11	7	72	-	6	37	-
	12	1	19	-	1	14	-
β	6	823	27,949	$3.19E^{-188}$	752	22,983	$8.06E^{-051}$
	7	924	7,443	$1.33E^{-005}$	470	6,174	$9.26E^{-001}$
	8	212	2,085	$4.38E^{-003}$	130	1,660	-
	9	46	724	$5.07E^{-001}$	26	427	-
	10	13	140	-	9	104	-
	11	1	32	-	5	27	-
	12	-	-	-	1	5	-
γ	6	823	26,694	$3.93E^{-003}$	780	23,016	$1.65E^{-001}$
	7	181	7,502	$3.55E^{-001}$	184	6,182	-
	8	54	1,683	-	57	1,668	-
	9	6	403	-	15	432	-
	10	2	131	-	2	128	-

Table 3.5: SEED (MATCH) TYPES. The method stops in case the proportions of functional (Fun.) and non-functional sites of seed match type of length (Len.) $l = a$ and seed match types of length $l > a$ are not significantly different (two-tailed Fisher text, P-value < 0.05). This procedure is applied separately for each start position ($\alpha = 1$, $\beta = 2$, $\gamma = 3$) [1].

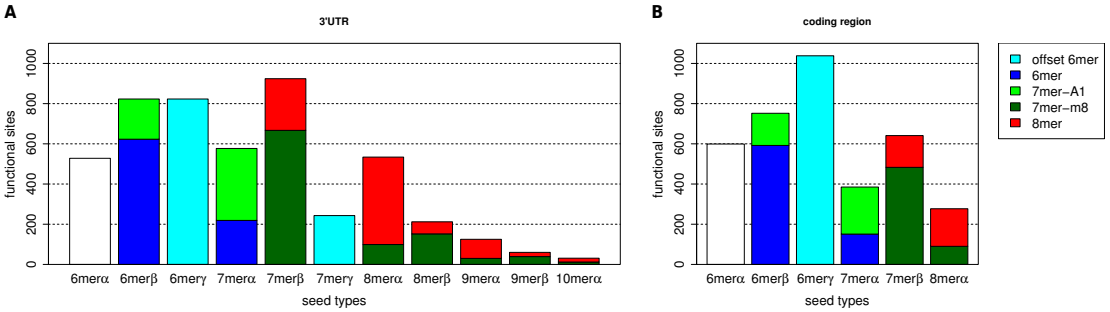


Figure 3.18: Overlap between seed type sets. The bars indicate the numbers of functional seed matches of the seed types described in this work for 3'UTR (A) and coding region (B). The coloring of the bars represents the partitioning of the seed matches on the classical seed types proposed by Bartel and colleagues [7]. Importantly, all seed matches complementary to classical seed types were covered by the seed types defined in this work, whereas the reverse was not true.

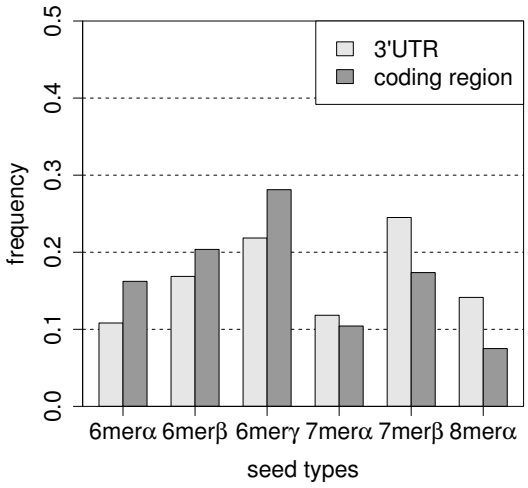


Figure 3.19: Comparison of seed type distributions. To facilitate comparison, seed types found for the 3'UTR but not for the coding region have been merged depending on the start position to 8merα, 7merβ and 6merγ, respectively.

	3'UTR		coding region	
Seed type	Sequences	CCRs	Sequences	CCRs
6merα	0.63	0.44	0.55	0.40
6merβ	0.81	0.70	0.77	0.65
6merγ	0.86	0.75	0.67	0.58
7merα	0.89	0.76	0.82	0.73
7merβ	0.87	0.78	0.82	0.74
7merγ	0.82	0.67	-	-
8merα	0.83	0.75	0.90	0.80
8merβ	0.89	0.80	-	-
9merα	0.78	0.63	-	-
9merβ	0.92	0.78	-	-
10merα	0.88	0.61	-	-

Table 3.6: Co-occurrence of seed types. For this statistic exclusively functional seed matches were considered. "Sequences" denotes the proportion of target sequences of a seed type that included matches of other seed types as well, e.g. 63% of the 3'UTRs containing a 6merα match included additionally matches of other seed types. "CCRs": analogous to "Sequences".

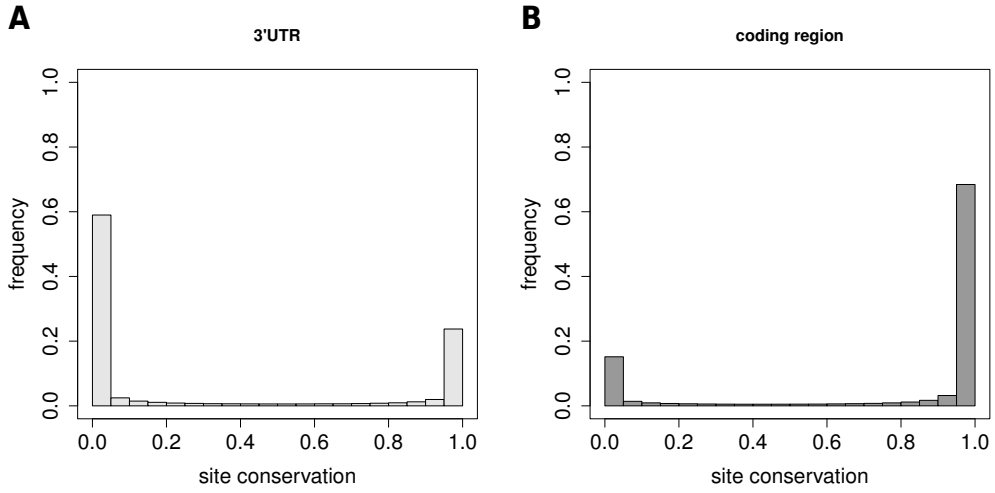


Figure 3.20: Background conservation. The histograms show the frequencies of conservation scores assigned to the sites of 3'UTR background (A) and coding region (B) background.

	Sensitivity		Specificity		Precision		MCC	
Seed type	all	cons.	all	cons.	all	cons.	all	cons.
6merα	0.11	0.06	0.70	0.91	0.01	0.02	-0.08	-0.02
6merβ	0.17	0.09	0.78	0.94	0.03	0.05	-0.02	0.02
6merγ	0.17	0.10	0.79	0.94	0.03	0.06	-0.02	0.04
7merα	0.12	0.07	0.91	0.97	0.05	0.08	0.02	0.04
7merβ	0.19	0.12	0.94	0.98	0.11	0.18	0.10	0.12
7merγ	0.05	0.03	0.92	0.98	0.02	0.04	-0.02	0.00
8merα	0.11	0.07	0.98	0.99	0.14	0.20	0.10	0.10
8merβ	0.04	0.02	0.98	0.99	0.09	0.14	0.04	0.04
9merα	0.03	0.02	0.99	1.00	0.11	0.14	0.04	0.04
9merβ	0.01	0.01	0.99	1.00	0.06	0.13	0.01	0.02
10merα	0.01	0.00	1.00	1.00	0.06	0.11	0.01	0.01

Table 3.7: Predictive performance of 3'UTR seed types. Classification performance of seed types based on all seed matches or only the conserved (cons.) ones.

	Sensitivity		Specificity		Precision		MCC	
Seed type	all	cons.	all	cons.	all	cons.	all	cons.
6merα	0.16	0.13	0.71	0.78	0.02	0.02	-0.05	-0.04
6merβ	0.20	0.17	0.78	0.83	0.03	0.03	-0.01	0.00
6merγ	0.28	0.24	0.70	0.77	0.03	0.04	-0.01	0.00
7merα	0.10	0.08	0.92	0.94	0.04	0.04	0.01	0.01
7merβ	0.17	0.14	0.92	0.94	0.07	0.08	0.06	0.06
8merα	0.08	0.06	0.97	0.97	0.07	0.08	0.04	0.04

Table 3.8: Predictive performance of ORF seed types. Analogous to table 3.7.

	3'UTR		coding region	
Seed type	OR	P-value	OR	P-value
6merα	2.23	0	1.23	0.0457
6merβ	2.45	0	1.61	0.0030
6merγ	3.17	0	1.51	0.0009
7merα	3.05	0	1.14	0.1446
7merβ	4.01	0	1.40	0.0005
7merγ	1.96	0.0031	-	-
8merα	4.48	0	1.16	0.4069
8merβ	2.80	0	-	-
9merα	3.55	0	-	-
9merβ	3.32	0.0050	-	-
10merα	4.20	0.1534	-	-

Table 3.9: Conservation of seed matches. Functional seed matches were categorized according to seed types determined for 3'UTR and coding region. The OR values were calculated according to equation 3.5. These values are illustrated by the red points in figures 3.15 and 3.17, respectively. The P-values were computed as described in section 3.3.2.

	3'UTR			coding region	
Seed type	A	B	C	D	E
6merα	0	0.0009	0.1739	0.2753	0.9411
6merβ	0.0067	0.0005	0	0.9187	0.0471
6merγ	0.0202	0	0	0.9979	0.0033
7merα	0	0.0003	0	0.8137	0.9654
7merβ	0.0002	0	0	0.8947	0.2421
8merα	0	0	0	0.4470	0.9865

Table 3.10: Conservation of seed matches of miRNA clusters. The P-values were computed as described in section 3.3.2.

4 Evaluation of miRNA target site prediction methods

A lot of evaluations of miRNA target prediction tools have been published so far, e.g. [44],[45],[127],[128],[129]. Moreover, most releases of new methods or of updates include a comparison of different methods. Various types of data sets have served as reference data for performance evaluations:

- **mRNA expression** data from miRNA transfection or depletion experiments. mRNAs whose expression is affected by a (stimulated) change of miRNA concentration are considered as miRNA targets. With knowledge of the miRNA it can be checked how well the predictions correspond to the measured change of mRNA expression. A drawback is that the affected mRNAs can also be indirect targets.

- By the same principle also **protein expression** measurements were used for evaluation, see [44],[45].

- Further, **CLIP data**, see section 3.1.7, were applied to assess accuracy of prediction methods, see e.g. [126],[130],[131]. CLIP methods have the advantage to uncover only direct miRNA targets and additionally to locate the site of interaction within the target sequence, but on the other hand they do not measure the strength of repression.

The aim of this evaluation was to asses the ability of prediction methods to identify experimentally determined target sites within 3'UTRs and coding regions. Performing an evaluation on *target site level* restricts the choice of methods, see definition in figure 4.1. Most of the tools do not provide the location of miRNA target sites, but predict miRNA-gene interactions (*target level*). Six methods were assessed: Diana-microT [132],[105], EIMMo [133],[107], miRmap [131], mirSVR [130], PITA [134] and TargetScan [135]. Apart from the recently published miRmap, these are popular and frequently cited target site prediction methods. As benchmark the processed PAR-CLIP data set was used, see sections 3.3.1 and 4.2.1. In this work particularly the sensitivity of the methods has been

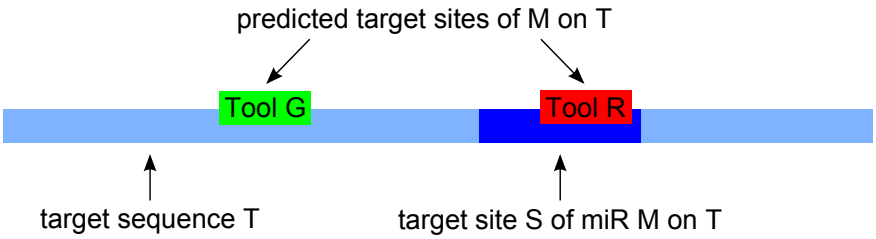


Figure 4.1: LEVELS OF EVALUATION. miRNA target prediction tools can be evaluated on two levels of different granularity, in this thesis referred to as *target level* and *target site level*. On target level the predictions of both tools, G and R, are true positives as both correctly identify the interaction between target T and miRNA M. On target site level the prediction of method G is wrong, i.e. it is a false positive, as it does not identify a true target site. Further, as method G does not detect target site S, S is counted as false negative with respect to method G. The prediction of tool R identifies a true target site of M and therefore is classified as a true positive on target site level.

investigated. To facilitate comparison between 3'UTR and coding region, only seed types identified in both 3'UTR and coding region have been considered, see figure 3.19. 3'UTR-unique seed types were merged with 6merγ, 7merβ and 8merα, respectively. According to the specifications of the tools, the focus is clearly on long seed matches, see [1]. But in section 3.4.1, it was demonstrated that 6mer seed matches cover the bulk of functional target sites. Further, in the light of the ceRNA hypothesis allowing short but less potent seed sites to be conclusively explained, see section 3.5, and if comprehensive elucidation of miRNA biology is sought, one may not disregard more than half of the functional sites. In the second part of this chapter, the accuracy of the methods was analyzed. This analysis continued a study presented on the German Conference on Bioinformatics, 2010, [125].

4.1 Background: Computational target site prediction

Due to the large numbers of miRNAs and their potential targets and the condition dependence of miRNA-target interactions, a comprehensive experimental determination of the miRNA-target interactome is practically not feasible [129],[136]. Albeit the currently estimated extent of miRNA-mediated regulation could not be anticipated at the beginning of the era of miRNA research, very early, first computational methods have been developed for the prediction of miRNA targets.

In section 3.1.2, the development history of TargetScan has been extensively illuminated. It was one of the first methods next to miRanda [137] and the approach developed by Stark et al. [138]. Meanwhile, there are dozens of methods, each with a special focus on the miRNA-target interaction. The following presents a survey of miRNA target prediction. A selection of prediction tools is described in more detail. These methods were evaluated in this thesis. Beforehand, various kinds of target site features commonly adopted by target prediction programs are outlined.

4.1.1 Categories of target site features

A seed match represents the clearest indications for a target site. Almost all target prediction tools perform a search for seed-complementary sites on the target sequence [136]. Many of them require perfect complementarity between seed and target. But some methods tolerate single mismatches or wobble pairing in the seed match as well. Most of them consider more than one of the seed types displayed in figure 3.2. Long seed types are preferentially used as they ensure higher specificity. But apart from the seed rule, a couple of further target site parameters is taken into account as the exclusive seed match would be to unspecific. With a length of 6 or 7 nt, many of the seed matches can occur simply by chance. Therefore, recording seed matches is commonly only the first step of target prediction. It creates the initial set of target candidates that will be filtered or prioritized in subsequent steps.

Evolutionary conservation Bartel emphasized in his review that "the use of preferential evolutionary conservation was a key methodological advance" to distinguish miRNA target sites from the 3'UTR background [7]. Analyzing orthologous sequences from multiple species allows for the identification of conserved sequence elements. Opposed to other sequence regions, conserved elements are protected by purifying selection against deleterious modifications. If a site is more conserved than expected by chance, it is most likely biologically functional. Consequently, the percentage of true positives can be increased by filtering conserved seed matches. It is important to point out that filtering conserved sites means neglecting non-conserved functional sites. Different approaches are used to judge conservation of target sites. Some require conservation of the seed match at orthologous sites, e.g. TargetScan [86] and miRmap [131], miRanda [120]. For this, the orthologous sequences are aligned and the seed match has to be within an aligned block. Less strict methods merely demand the presence of a seed match for the miRNA (somewhere) in each of the orthologous sequences, e.g. mirWIP [139] and rna22 [140].

By the use of PhastCons scores, the alignment-based variant was applied in this work.

Co-evolution of target and miRNA This approach is based on the assumption that biologically functional sites in a sequence such as miRNA seed matches differ in their composition from the remaining sequence. In contrast to non-functional regions, functional miRNA seed matches are protected by purifying selection against mutations that would adversely affect the miRNA-target interaction. That is, instead evolving like the surrounding sequence, the seed match evolved synchronously with the matching miRNA. For instance, Marin et al. implemented this approach as follows [141]: A 3'UTR includes n seed-complementary sites for a miRNA. By using a Markov model representing the composition of the 3'UTR, the probability is computed to find n sites of same length by chance in the 3'UTR. The lower the probability is, the higher the chances that the miRNA-3'UTR interaction is functional.

Sequence features beyond the seed match Apart from the seed match, the target site includes further sequence-based features that contribute to the specificity of the miRNA-target interaction. There are two types of pairing involving the remainder of the miRNA, *supplementary* and *compensatory* pairing. Supplementary target sites contain beside a perfect seed match additional pairing between the target and the 3' portion of the miRNA. Grimson et al. observed 3' pairing to be efficient if miRNA positions 13-16 matched contiguously to the target [94]. Beside supplementing, 3' pairing was found to compensate for mismatches or bulges within the seed match. But both compensatory and supplementary sites seem to be rare [7]. A more common feature is the local base content of functional seed matches [7]. Grimson et al. identified a bias to adenines and uracils in immediate proximity to functional sites [94]. In particular, conserved sites were flanked by a high AU content. Importantly, the AU enrichment dropped quickly with increasing distance to the site. Another context feature is the position of the functional site within the 3'UTR. Both Majoros et al. [142] and Grimson et al. [94] observed that seed matches were not evenly distributed along the 3'UTR. Functional sites aggregated near the two ends of the 3'UTR, stop codon and poly(A) site. At the stop codon end the distribution peak was offset approximately 15 nt downstream of the stop codon. Grimson et al. could show that target sites located close to the 3'UTR ends were both more effective in target destabilization and more conserved than sites occurring centrally.

Thermodynamic stability of the miRNA-target interaction The interaction between miRNA and target can be considered from a thermodynamic perspective. One hypothesis is that functional interactions are based on stable miRNA-target duplexes. Here, the hybridization of the entire miRNA with the target is taken into account. The more stable the RISC is bound to the target via the miRNA, the more time is available to carry out the regulation [136]. Alternatively, a stable hybridization causes a higher net gain of energy than a unstable one and thus it is energetically more favored. The net gain, represented by the *free energy* ΔG_{duplex}, is the difference between the energy needed for the creation of the duplex and the energy obtained by the hybridization. It is important to note that thermodynamic stability is not well compatible with the enriched AU content observed in the vicinity of functional seed matches. Base-pairing between G and C is more stable than pairing between A and U. Consequently, if the seed match-flanking regions have an increased GC content the interaction will be predicted to be more stable.

Structural accessibility of the target site mRNAs do not exist as linear molecules in the cell but form spatial structures. RNA structures due to intra-molecular base-pairings are called secondary structures. Secondary structures can influence the accessibility of target sites, e.g. Kertesz et al. showed experimentally that the efficiency of miRNA-mediated regulation could be significantly reduced through mutations diminishing target site accessibility [134]. Target sites are preferentially located in structurally accessible regions [136]. If not, structural transformations have to precede miRNA-target interaction. The accessibility of mRNA regions or respectively the energetic costs to open hampering secondary structures can be quantified with secondary structure prediction algorithms. For instance, the PITA algorithm, see section 4.1.2, combines the energy costs needed for unwinding the target site and flanking regions with the free energy obtained from the subsequent hybridization between miRNA and target [134]. Since A and U form less stable base-pairs, the local AU content and the accessibility of target sites correlate with each other [7],[131].

4.1.2 Survey of prediction algorithms

Diana-microT The Diana tools have been developed by the Hatzigeorgiou lab and comprise beside others an algorithm for predicting target sites in both 3'UTR and ORF, microT-CDS [105]. microT-ANN is a method exclusively designed for the prediction of 3'UTR target sites [132]. Both methods start with compiling a set of putative target sites. The highest scoring alignment is determined for each 9 nt long sequence fragment of the target sequence and the miRNA *driver sequence*,

that is, the first 9 nt of the miRNA. Putative target sites need to have at least 6 Watson-Crick (WC) pairs between driver and target. Further, a consecutive match of 4 WC matches [105], respectively of 6 matches (WC or G:U wobble) [132] is required to start either at position 1 or 2 relative to the miRNA. Position 3 is not considered as a start position. If the alignment of driver and target contains 6 WC pairs, a single wobble pair is allowed. microT-CDS allows a single mismatch or bulge for target sites with 8 WC matches.

Putative sites are filtered depending on the free energy. Depending on length, start position, 3' pairing and presence of an A opposite to the first miRNA position, target sites are grouped to *binding categories*. Sites with less than 7 WC bps are removed if the free energy between target and driver does not exceed the threshold for the respective binding category. Conservation scores are determined for the filtered target sites. This score represents how often a site could be identified at the exact same position in a set of orthologous 3'UTRs. The final score of a target site depends on the binding category, the miRNA and the conservation score: For each binding category and each conservation score the number of target sites of a certain miRNA is compared with the number of sites of corresponding mock miRNAs having an equal or greater conservation score. The ratio of the two counts defines the target site score.

Reczko et al. improved the prediction of miRNA-gene interactions by considering additionally target sites located in the coding region [105]. They created sets of true sites and false sites similarly to the approach used in this work, see section 3.3.1. Using logistic regression between the binding categories of aligned sites, 64 in total, and the presence or absence of the sites in the true or false set, they found 21 significant binding categories for sites in the 3'UTR and 10 categories for sites in the ORF. Beside the weight obtained for a significant binding category (i.e. the regression coefficient) they considered following target sites features: conservation, local AU content, distance to 3'UTR or ORF ends, distance between adjacent target sites, free energy, accessibility and the binding pattern between target site and miRNA. Generalized linear models were used to integrate the features and to compute the target site score. ORF and 3'UTR are represented by separate models.

EIMMo Gaidatzis et al. did not refer to the set of classical seed types [133]. They analyzed the conservation of nine hypothetical seed types. Some of these included either a single mismatch, bulge or loop. 6mer and imperfect seed types differed only slightly from the background conservation. Further, requiring an adenine in the target opposite to miRNA position 1 as Lewis et al. [86] entailed no clear

enrichment of conserved sites in their analysis. Therefore, EIMMo considers only sites with perfect complementarity to miRNA sub-sequences 1-7, 2-8 and 1-8.

EIMMo starts with a search for candidates sites and then scores the conservation of the putative target sites based on a Bayesian approach. Given the conservation of a seed match in different species, the posterior probability is calculated that the site is functional. The phylogenetic relationships of the considered species flows into the score calculation. Another unique feature is that EIMMo creates miRNA-specific selection models of target sites. Thus, the possibility is taken into account that the selective pressure to maintain target sites in related species could be different between miRNAs.

miRmap Vejnar et al. created a comprehensive open-source library that covers all feature categories presented above [131]. A total of 11 different target sites features were implemented, three of which were new. In a detailed analysis, they examined the correlation of all feature pairs and they evaluated the predictive power of each feature based on various experimental data sets. Based on that Vejnar et al. developed the target site prediction method miRmap combining a total of 10 features from the five feature categories. The features were combined by multiple linear regression optimizing the prediction of the target expression change.

mirSVR mirSVR [130] is an extension of the miRanda algorithm [120]. miRanda searches for maximal local complementarity alignments between the target and the complete miRNA sequence. That is, it differs to approaches scanning the target merely for seed matches. But miRanda accounts also for the importance of the seed in target recognition by assigning higher position-specific weights to matches involving the seed region. mirSVR obtains the miRNA-target duplexes from miRanda. These target site candidates may include a single G:U wobble or mismatch within miRNA positions 2 to 7. Except co-evolution of miRNA and target, all feature categories mentioned in section 4.1.1 are considered in the determination of the score by mirSVR. To learn the feature weights, mirSVR was trained on miRNA transfections experiments using support vector regression.

PITA Kertesz et al. showed that it is necessary to consider target site accessibility in order to determine the efficacy of target repression by miRNAs [134]. Further, a genome-wide analysis revealed that positioning of target sites in accessible regions is a conserved feature in genomes. An energy-based score, $\Delta\Delta G$, was conceived representing the difference between the energy needed to unpair the

target region, ΔG_{open}, and the free energy gained by binding of the miRNA to the target, ΔG_{duplex}, see above. That is, $\Delta\Delta G = \Delta G_{open} - \Delta G_{duplex}$. $\Delta\Delta G$ strongly correlated with the degree of target repression whereas exclusive ΔG_{duplex} showed only poor correlation in their analysis. The miRNA is attached to a large protein complex, RISC, that is also interacting with the target. Consistent with this, the correlation between $\Delta\Delta G$ and target expression could be appreciably improved when taking into account the regions next to the target site in the calculation of ΔG_{open}. Subsequent to a search for seed matches, the PITA algorithm computes the $\Delta\Delta G$ for each putative target site.

TargetScan The first version of TargtScan was published in 2003, i.e. meanwhile TargetScan can look back upon 10 years of development history. The current TargetScan algorithm, version 6.2, builds on TargetScanS that has been released in 2005 [86]. TargetScanS searched for conserved sites perfectly matching to seed types 7mer-A1, 7mer-m8 and 8mer, see figure 3.2. The site had to occur at corresponding positions in a multiple alignment (i.e. within an aligned block) of orthologous 3'UTRs from five vertebrates to be considered conserved. In 2007, Grimson et al. described five sequence-context features beyond seed pairing involved in target recognition [94]. Three of which, the local AU content in immediate proximity to the target site, pairing of the 3' portion of the miRNA and positioning within the 3'UTR were integrated as *context score* into TargetScan. The relationships between the individual context features and target downregulation were determined by linear regression. Importantly, the enhancement by the context features even allowed for prediction of effective non-conserved sites. Friedman et al. overhauled the method to evaluate site conservation in 2009 and introduced a new criterion: P_{CT}, the probability of preferentially conserved targeting allows for assessing the biological relevance of predicted miRNA-target interactions complementary to the context score [87]. In 2011, Garcia et al. found that varying targeting proficiency between miRNAs is connected with both the seed-pairing stability (SPS) and the target-site abundance (TA) of miRNAs [135]. The proficiency of miRNAs was integrated in the context-score model of TargetScan, now referred to as *context+ score*.

4.2 Methods and Materials

4.2.1 Preparation of prediction databases

Six target site prediction programs were evaluated in detail, Diana-microT [132], [105], EIMMo [133],[107], miRmap [131], mirSVR [130], PITA [134] and TargetScan [135]. Developers of target prediction methods usually provide both pre-computed genome-wide predictions and an executable program to predict targets on custom data. The prediction databases were intersected with the sets of functional and non-functional sites for the 3'UTR or respectively the coding region (in the following referred to as *benchmark data sets*) that are introduced in section 3.3.1. The initial version of the 3'UTR benchmark data set comprised 1,873 3'UTRs and 38 miRNAs. Overall 4,880 functional and 129,509 non-functional sites were distributed on the 3'UTR sequences. The latter number is slightly reduced compared to that presented in section 3.3.1, since sites for which no PhastCons score was available were not considered for the evaluation. The initial ORF benchmark set included 3,692 functional and 104,845 non-functional sites of 38 miRNAs located on 1,580 coding regions.

Predictions that did not match exactly to one of the benchmark sites were ignored. Removed sites were either located too close to the borders of the respective regions or the seed match was imperfect, i.e. it contained mismatches or wobble pairing. Most of the examined methods do not search explicitly for seed matches starting at position 1 relative to the miRNA. Nevertheless, they are able to identify 7merα and 8merα sites as 6mer and 7mer matches starting at position two are covered by the former two.

In case of TargetScan, mirSVR, miRmap, EIMMo and Diana-microT pre-computed predictions were used as these databases are regularly updated. The PITA database was updated last time in 2008. Thus, the program was run to predict target sites on the sequences of the benchmark data set. Subsequently, the preparation of the individual data sets for the evaluation is described.

Diana-microT The Diana lab tool webpages, http://62.217.127.8/DianaTools/, provide access to the Diana-microT predictions databases via a REST web service interface. Diana-microT-CDS, version 5, produces predictions for both 3'UTR and coding region. Of these, only the predictions for the ORF were considered. In terms of the 3'UTR, it was resorted to the predictions by the algorithm Diana-microT-ANN, version 4, that generates predictions exclusively for this region. Diana tools use sequence material from Ensembl. The retrieved predictions were based on Ensembl version 69 [143]. The corresponding RefSeq sequences were de-

termined by sequence comparison. Intersecting benchmark sites and microT-CDS predictions yielded 6,067 sites residing on 853 coding regions. With regard to the 3'UTR, 13,846 sites could be found on 571 sequences.

EIMMo The EIMMo target prediction database was updated the last time in January 2011 (http://www.mirz.unibas.ch/, version 5). Here, the positions of the predicted target sites are specified by their chromosomal coordinates based on human genome assembly hg19. In the first step, the coordinates were converted to coordinates of hg18 using the liftOver utility from UCSC [117]. For sites completely included in 3'UTRs of the benchmark data set, the positions relative to the 3'UTR were calculated by means of the chromosomal exon coordinates of the 3'UTRs. 28,525 sites were located on 1,832 3'UTRs of the benchmark set. 28,206 of the predictions matched with benchmark sites. The predictions for the coding regions were analogously prepared. 21,303 of 21,568 predictions distributed on 1,500 coding regions overlapped exactly with benchmark sites.

miRmap The most recent miRmap version was from January 2013. The target sequences used come from Ensembl, version 69, [143] and could be downloaded from the miRmap website, http://mirmap.ezlab.org/. Before the predictions could be intersected with benchmark sites, the Ensembl 3'UTRs had to be mapped to the RefSeq 3'UTRs from the benchmark data set. By sequence comparison, it was found that 749 benchmark sequences containing 23,743 potential target sites were exactly included in the miRmap target set. 23,716 of these sites were contained in the benchmark set.

mirSVR mirSVR was released in August 2010. Four files are provided for download on http://microrna.org/. Predictions for conserved and non-conserved miRNAs are separated. These two sets are further subdivided into highly scored and lowly scored potential target sites, respectively. The positions of the target sites are not directly specified in the downloaded files, but the alignments of the miRNAs and the target site regions are provided. The location of the alignments are indicated by their chromosomal positions relative to genome assembly hg19. Using the liftOver program [117], the locations were converted to coordinates of hg18. The chromosomal coordinates of the longest perfect seed match were deduced from the alignment representation. A match had to be at least 6 nt long and had to start at position two or three relative to the miRNA. By means of the chromosomal exon coordinates, the positions of the predictions relative to the 3'UTRs of the benchmark data set were calculated. In this way 55,295 seed matches could

be assigned to 1,856 3'UTRs. Intersecting the mirSVR sites with the benchmark sites yielded 54,928 predictions that are distributed on 1,854 3'UTRs.

PITA PITA (http://genie.weizmann.ac.il/pubs/mir07/) was run with default parameter settings on the sequences of the 3'UTR benchmark set. The PITA algorithm seeks for seed matches of length 6, 7 and 8 starting at position 2 relative to the miRNA. For matches of length 7 and 8, 1 wobble pairing is allowed per site. Within matches of length 8, 1 mismatch is permitted additionally. 322,425 potential target sites were found by PITA on the 1,873 benchmark 3'UTRs. After intersecting with the benchmark sites, the PITA prediction set consisted of 72,399 sites located on 1,868 sequences.

TargetScan The TargetScan predictions were based on TargetScan version 6.2 which has been released in June 2012 (http://www.targetscan.org/). Conserved and non-conserved seed sites are listed in separate files. Beside target site predictions also the target sequences are provided for download. The sequences included in TargetScan Release 6.2 were based on human genome assembly hg19. Since transcript sequences may vary between different assemblies, the sequences used for the TargetScan prediction and the benchmark sequences were intersected. 1,691 non-redundant sequences of the TargetScan set showed complete identity to one of the 3'UTR sequences of the benchmark set. 26,154 sites located on 1,666 sequences matched with sites of the benchmark data set.

Not all sequences or miRNAs of the benchmark data set were included in each of the prediction databases. Therefore, after all predictions had been processed and intersected with the benchmark data set, the benchmark data was revised. It was adjusted in order to contain only those sequences and miRNAs that were included in each of the databases. Hence, the final benchmark data set for the 3'UTR consisted of 537 3'UTRs, 38 miRNAs, 1,401 functional and 40,982 non-functional sites. The ORF data set included 913 coding regions, 38 miRNAs and 2,116 functional and 57,544 non-functional seed matches. The respective numbers of sites per tool are shown in table A4.3[1].

4.2.2 Evaluation measures

Functional sites as defined in section 3.3.1 were considered as *observed positives* (OP). Analogously, non-functional sites were categorized as *observed negatives*.

[1]The numbers of figures or tables included in the Appendix are preceded by an "A".

Predictions considered in this evaluation were either functional or non-functional. Thus, predictions overlapping with observed positives are termed *true positives* (TP) and the intersection of predictions and observed negatives includes *false positives* (FP).

Following measures have been used to assess the performance of the prediction methods. *Sensitivity* indicates the proportion of identified observed positives. *Precision* represents the proportion of true positives included in a set of predictions. Sensitivity and precision characterize complementary properties of a classification system. To assess holistically the quality of a method, the *F-measure* was calculated which is the harmonic mean of sensitivity and precision.

$$Sensitivity = \frac{TP}{OP} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

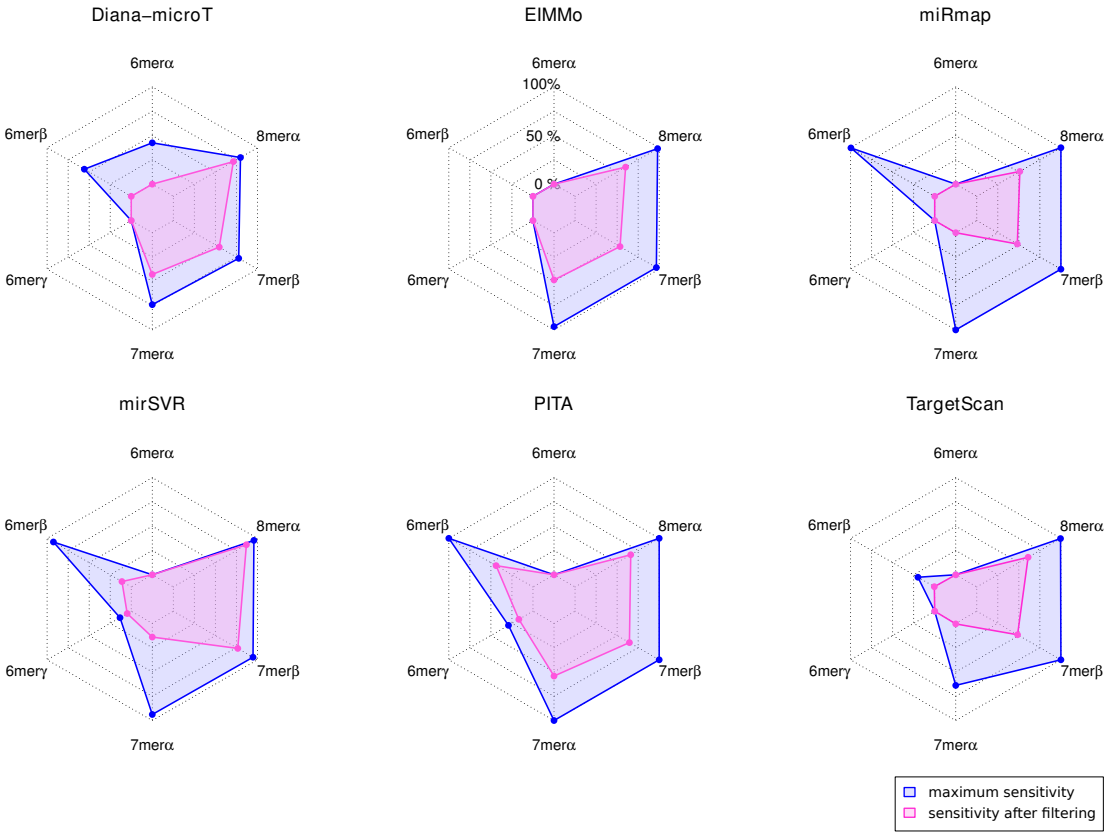
$$F\text{-}measure = 2 \cdot \frac{Sensitivity \cdot Precision}{Sensitivity + Precision} \tag{4.3}$$

4.3 Results

4.3.1 Sensitivity of prediction methods

Figures 4.2 and 4.3 show for each prediction tool the proportion of functional seed matches per seed type that have been predicted. Two sensitivity values were calculated. First, the maximum achievable sensitivity was determined per seed type by ignoring the target site scores. Alternatively, as this sensitivity is the result of the initial seed search, it was also called *initial* sensitivity in this work. The maximum or initial sensitivity is represented by the light blue colored proportions. Further, for each tool the score value t was determined that achieved the best F-measure on this benchmark data set. The magenta-colored proportions indicate the sensitivity after predictions with a score worse than t had been removed.

Sensitivity of 3'UTR predictions As only method Diana-microT discovered 6merα sites in the 3'UTR, see figure 4.2. According to its specifications, sites starting at position 3, i.e. position γ, did not belong to the set of predictions. Before scores are calculated, Diana-microT filters putative target sites depending on the free energy. Therefore, for none of the seed types, 100% of functional sites were found. Filtering sites with a score worse than t removed mostly less specific short seed matches. The number of functional 8merα sites was only marginally



.

Figure 4.2: Sensitivity of 3'UTR predictions per seed type. The radar plots display for each method the maximum obtainable sensitivity per seed type. Further, the sensitivity values after sites with a target site score $< t$ have been removed is highlighted. The method-specific cut-off values t correspond to the score achieving the highest F-measure.

diminished. EIMMo searches for sites of length 7 and 8 and disregards 6mers completely. Filtering reduced the sensitivity of all three seed type almost equally. Again, 8merα sites were relatively least affected. The miRmap prediction database contained all seed matches according to the classical seed types except the offset 6mer or respectively 6merγ. None of the functional 6merβ and 7merα sites exceeded the score cutoff. These sites dropped completely out of the prediction set. On the other hand, about 50% of both 8merα and 7merβ sites were retained. mirSVR requires the alignment between miRNA seed and target to include miRNA positions 2-7. Therefore, target sites including a 6merα seed match were not found. As mirSVR allows mismatches within the seed alignment to a limited extent, the

fraction of 6merγ sites that are included in longer imperfect matches could be detected. 6merβ and 7merα sites were strongly affected by the filtering. Similar to mirSVR, PITA allows mismatches in sites of length > 6 nt. Start position 1 is not considered. Here, filtering affected all types of seed matches similarly. TargetScan searches for following subset of the classical seed types: 7mer-A1, 7mer-m8 and 8mer sites. To discover 7mer-A1 sites, TargetScan scans for 2-7 matches preceded by an adenine on the target. If the matching miRNA starts with an uracile, the site is of type 7merα otherwise it is of type 6merβ. As the majority of miRNAs starts with an uracil, significantly more 7merα matches were included in the initial TargetScan prediction set. Similarly to miRmap and mirSVR, 6merβ and 7merβ seed matches did not achieve sufficiently high scores and thus were removed by the filtering.

As described already in [1], a common trend to the long seed matches 7merα, 7merβ and 8merα was clearly visible. Sites of these types were preferentially searched and scoring functions assigned the highest scores to them. Regarding 6mer sites, 6merβ sites were searched by five of six methods. 6merγ sites were not explicitly regarded by any of the methods. However, fractions of 6merγ sites were detected by tools that allow for a mismatch or G:U wobble within the seed match. A fraction of 6merα sites was discovered by Diana-microT.

Sensitivity of ORF predictions The study of miRNA-target interaction through ORF sites has recently begun. For this reason, not many methods do exist for predicting target sites specifically for this region. Subsequently, EIMMo and Diana-microT were analyzed.

The ORF-located sites predicted by EIMMo were generated by the algorithm that has been originally designed for the 3'UTR [133]. Only the sequence data was exchanged for the prediction in the coding region [107]. Consequently, EIMMo predicted the same types of sites as for the 3'UTR. Further, filtering affected sites of all types similarly.

In contrast, Diana-microT, more precisely Diana-microT-CDS, is a new approach estimating possible miRNA-target relations by considering both 3'UTR sites and ORF sites [105]. To understand better the performance of Diana-microT, the seed model used by Reczko et al. has to be envisioned first. As mentioned in section 4.1, Reczko et al. defined binding categories [105] and checked each of which for significant enrichment in CCRs using the data of Hafner et al. [9]. As for the 3'UTR, they took only start positions 1 and 2 into account. Further, it was distinguished whether the base opposite to the first miRNA position was an adenine or not and whether the 3' end of the miRNA was involved in pairing or not. In total,

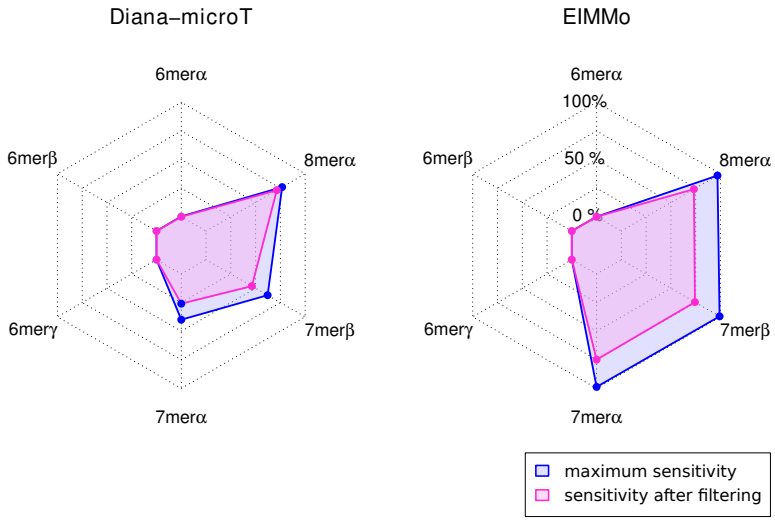


Figure 4.3: Sensitivity of ORF predictions per seed type. Captions analogous to figure 4.2.

they defined 64 different binding categories, e.g. 7merα sites are distributed on four binding categories. Of these only category '7mer 1st:match+A', representing sites with an A opposite to the first miRNA position and no additional 3' pairing, turned out to be significant. Overall 10 categories were significantly represented with respect to the coding region. Surprisingly, neither of the 6mer categories was among them. Unfortunately, a complete listing of all binding categories considered is missing in [105]. From a length of 8, it was not clear from the description which categories were actually considered.

Diana-microT-CDS preferentially identified functional 8merα seed matches, followed by 7merβ sites. Less than 50% of the functional 7merα sites were included in the initial set. 100% sensitivity was not achieved for any of the three seed types due to the division of sites of one seed type on several binding categories of which only a fraction was significant. While the fractions of functional 7mer sites were reduced through filtering, the number of 8mer sites remained almost the same.

4.3.2 Differences between scoring systems

Prediction methods include scoring systems to add a value indicating the trustworthiness of the prediction to each predicted target site. In the following, the scoring systems or functions of the predictions tools were analyzed based on their score value distributions. It was of particular interest whether and how the scoring systems distinguish between seed types. Figures A4.9 and A4.10 show the normalized

	homogeneous score distributions (3'UTR)			
Diana-microT	6merα, 6merβ	7merα, 7merβ	8merα	
EIMMo	7merα, 7merβ	8merα		
miRmap	6merβ, 7merα	7merβ, 8merα		
mirSVR	6merβ, 7merα	6merγ	7merβ	8merα
PITA	6merβ	6merγ - 8merα		
TargetScan	6merβ, 7merα	7merβ	8merα	
PhastCons	6merα - 7merβ	6merβ - 8merα		

Table 4.1: Distinction between seed types by the scoring systems (3'UTR). Scores assigned to functional target sites were grouped according to seed types. For each method each pair of seed type-related score value distributions was tested on homogeneity using the Wilcoxon test with Bonferroni correction. Seed types with homogeneous target site score distributions (P-value < 0.05) are co-placed in a cell of the table. The names of the seed types within a cell are ordered lexicographically.

score values of 3'UTR and ORF predictions, respectively. For comparison a naive approach, called "PhastCons", was analyzed additionally. It contains all functional and non-functional sites of the benchmark set. The PhastCons conservation scores were used as target site scores. How the filtering changes the sensitivity of the naive approach is displayed in figure A4.8.

To test if seed types were treated differently, all pairs of seed type-related score value distributions were tested on homogeneity, see tables 4.1 and 4.2. The reasons for the variations between score distributions are discussed separately for 3'UTR and coding region.

Scoring of 3'UTR predictions PITA and the naive approach had the most simple scoring systems. Both methods involved only one feature and the feature score is also the target site score, i.e. $\Delta\Delta G$ and the PhastCons score, respectively. The score value distributions of functional sites of PITA and the naive approach exhibited the least variation between seed types of all methods. The PITA scoring system merely rated 6merβ sites differently. Except for the difference between 6merα and 8merα sites, PhastCons scores did not significantly vary between seed types. Also Diana and EIMMo rated seed matches by means of their conservation. But both methods calculate the target site score based on a seed type-specific signal-to-background ratio. As illustrated by the score distributions of Diana-microT

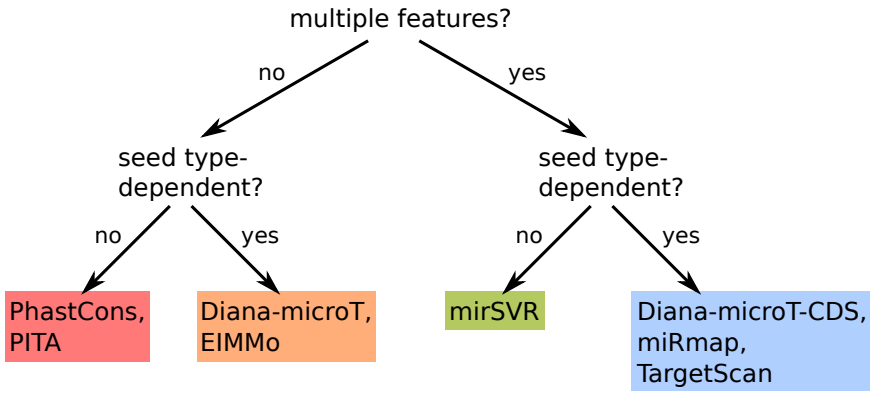


Figure 4.4: DIFFERENT SCORING SYSTEMS. The predictions methods were grouped based on their scoring systems. First, it was regarded if one or more features are implemented. The next subdivision depended on whether the seed type was considered for the calculation of the target site scores or not.

3'UTR predictions, taking into account the background conservation created a discrepancy between long and short seed types.

mirSVR, miRmap and TargetScan combine several features to assess the credibility of seed matches. Further, all three methods weighted the features according to their effect on target expression. A consequence that arises from correlating the score with repression strength was that the scores of 7merα sites were on the same level as the scores of 6mer sites, see table 4.1, since the impact on transcript stability mediated through 7merα sites is comparably weaker than through 7merβ and 8merα sites, see figure 3.9. Interestingly, the conservation-based methods such as Diana-microT and EIMMo rated the two 7 nt long seed types equally.

TargetScan and miRmap consider different sets of seed types. TargetScan distinguishes between seed types 7mer-A1, 7mer-m8 and 8mer that give rise to three different groups of score distributions, see table 4.1. The small proportion of 6merβ sites identified by TargetScan were scored the same as the 7merα seed matches. miRmap seeks explicitly for sites matching to positions 2-7 and 2-8 relative to the 5' end of the miRNA, that is 6merβ and 7merβ. But since matching to position 1 is not excluded, 7merα and 8merα were also identified in this way. The scoring system of miRmap, however, assessed 7merα and 8merα just as 6merβ or 7merβ, respectively. Unlike TargetScan and miRmap, mirSVR has a seed-type independent scoring system, i.e. the seed type is not taken into account for the score calculation. Nevertheless, the mirSVR scores showed the greatest variations between seed types, see figure A4.9. Strikingly, 7merβ and 8merα sites were much

	homogeneous score distributions (ORF)	
Diana-microT	7merα, 7merβ	8merα
EIMMo	7merα - 8merα	
PhastCons	6merα - 8merα	

Table 4.2: Distinction between seed types by the scoring systems (ORF). Caption analogous to table 4.1.

higher rated than the remaining sites.

Concluding, the differences between the scoring systems were due to the features implemented and whether the features were weighted dependent or independent of seed types. Figure 4.4 shows a classification of target site prediction methods based on these two aspects.

Scoring of ORF predictions The ORF predictions of both EIMMo and the naive approach were generated by the same algorithm as the 3'UTR predictions. The EIMMo scores of 8merα sites were not significantly different to the scores assigned to 7merα and 7merβ sites, see table 4.2 and figure A4.10. The lower score level of ORF-located 8merα sites is possibly to be attributed to the less importance of this seed type for the coding region as described above, see figure 3.19.

Examining the score value distributions of the naive method revealed a significant increase of the scores reflecting the high conservation of the coding region. The difference between the scores of 6merα and 8merα sites observed in the 3'UTR was not observable for sites located in the coding region. Further, consistent with the observations in section 3.4.3, both functional 7merα and 8merα sites were not significantly stronger conserved than the corresponding non-functional sites, see figure 4.10.

The ORF release of Diana-microT involves multiple features, see figure 4.4. Importantly, in contrast to the other methods incorporating multiple features, the features were not combined with linear regression on mRNA expression data. But they were weighted by means of their ability to distinct between true and false target sites. For this reason, the score value distributions of 7merα and 7merβ sites were homogeneous. Surprisingly, 8merα sites achieved a higher score level than the 7mer seed matches. This is in contradiction to the observations presented in 3.4.1 and the study by Marin et al. [126] arguing that 8merα sites are less important in the coding region than 7merβ sites.

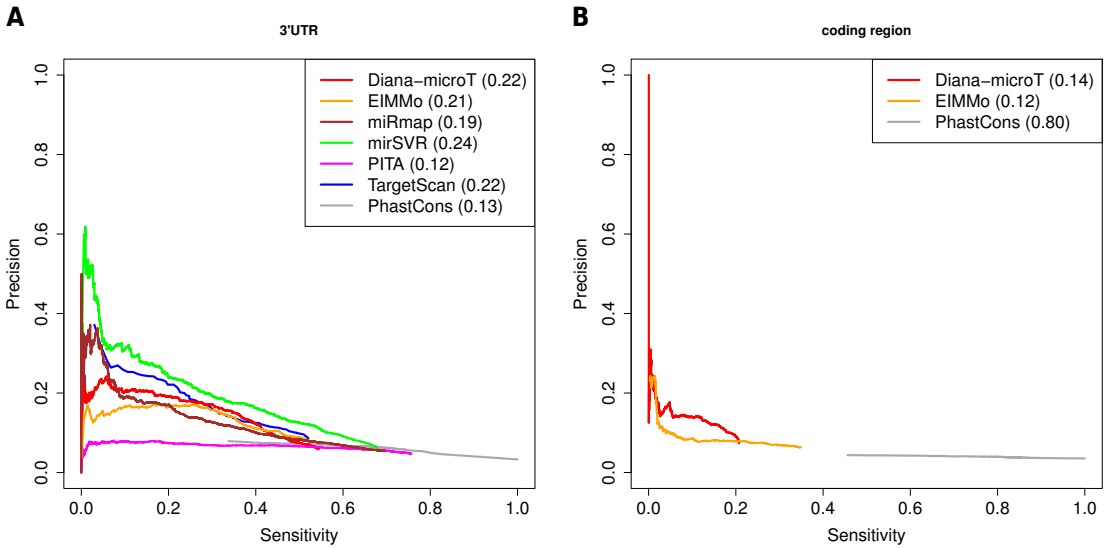


Figure 4.5: Precision-Sensitivity Curves. The values in brackets indicate the respective highest F-measure.

4.3.3 Accuracy of prediction methods

As the current models of miRNA target sites are insufficient, prediction tools are not able to distinguish precisely between true and false sites. For this reason, sensitivity can only be improved at the expense of precision and vice versa. Figure 4.2 shows that removing sites with smaller scores than the score related to the highest F-measure diminished sensitivity, e.g. the sensitivity of miRmap decreased from 0.69 to 0.29. On the other hand, the removal of low-rated sites increased precision (data not shown).

Here, the performance of the methods was assessed by means of the relationship between precision and sensitivity for all score values. Figure 4.5 displays the precision-sensitivity curves (PSC) for 3'UTR and ORF target site predictions. In general, it can be stated that the quality of miRNA target site prediction as was found here is low: Comparing the F-measure scores, mirSVR reached the highest F-measure (0.24) of all evaluated 3'UTR prediction methods and Diana-microT performed best with respect to the coding region (0.14).

First, the prediction methods for 3'UTR sequences were assessed. The rightmost point of a PSC represents the initial sensitivity (that is also the maximum sensitivity) and the initial precision of a prediction method. These values correspond to the initial set of seed matches. By successively raising the score cutoff and removal of sites with scores lower than the current cutoff, the sensitivity decreased and the precision increased. The extent of the precision gain depended on the quality of

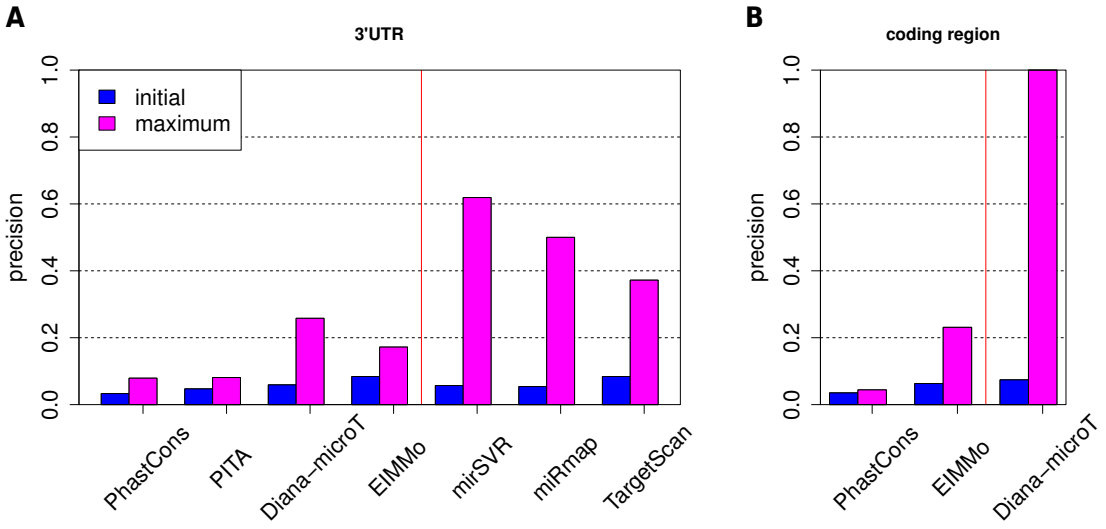


Figure 4.6: PRECISION GAIN. The meaning of *initial* and *maximum* precision are described in the text. The methods are arranged according to the grouping displayed in figure 4.4. Methods modeling only one target site feature (left side) are separated from methods involving several features (right side) by the red line.

the scoring system, see figure 4.6.

It turned out that methods with similarly configured scoring systems as shown in figure 4.4 exhibited similar PSCs. PITA and the naive approach performed worst, see figure 4.5(A). Compared with their respective initial precision values the precision increased by a factor of 1.7 and 2.4, respectively. The PSC of the naive approach stopped early since 34% of functional sites were associated with the highest PhastCons score. Due to the focus on more specific long seed matches, EIMMo had already a high initial precision. Through removal of sites with less credibility according to its scoring system the precision doubled. Diana-microT turned out to have the most discriminating scoring approach of the single-feature methods. Its precision improved by a factor of 4.4.

By contrast, the precision of mirSVR, the best performing tool of all, increased more than tenfold. Further, the precision values of miRmap and TargetScan grew by a factor of 9.3 and 4.4, respectively. Indeed, the relative gain of precision was similar for Diana-microT and TargetScan, but the initial set of TargetScan included more 7mer sites than that of Diana-microT (94% vs. 76%), see also 4.2. That is, Consequently, the precision of Diana-microT can be easily improved by removing unspecific 6mer seed matches, while TargetScan had to distinguish between true and false 7mer seed matches.

Basically, two types of PSCs appeared. Scoring systems of methods that involve multiple features such as mirSVR, miRmap and TargetScan generated a wide spectrum of precision values as opposed to prediction tools with only one feature (except Diana-microT). But also for the first three, in particular mirSVR and miRmap, the PSCs did not linearly increase with increasing cutoff. At about 10% sensitivity, these curves showed a sharp increase that probably represents the transition between short and long seed matches.

Considering the overall performance, mirSVR outperformed any other method. Its curve was constantly above the other PSCs. The ranking after mirSVR was less clear as the PSCs were crossing. For low sensitivity values miRmap and TargetScan showed high precision. But already for sensitivity values higher than 0.2 the miRmap PSC fell below the levels of Diana-microT and EIMMo whereas TargetScan remained longer above these two. PITA and the naive approach showed the worst performance. Interestingly, the latter slightly outperformed PITA in terms of the F-measure and the maximum gain of precision. Consequently, site conservation is better suited for miRNA target site prediction than site accessibility, at least in the 3'UTR.

Diana-microT was the only method particularly designed for target site prediction in the coding region. The EIMMo predictions were generated by the same algorithm as for the 3'UTR. Only the sequence material had been exchanged. The PSCs shown in figure 4.5(B) revealed Diana-microT to perform best. As long as it extended to the right, the PSC of Diana-microT was above the curve of EIMMo. Further, the scoring system achieved to increase the precision by a factor of 13.6 compared to the initial precision, even though the precision immediately dropped sharply. The PhastCons-based method was the worst of all three. It performed worse than in the 3'UTR as indicated by the F-measure and the precision gain. Consistent with previous observations this shows that discrimination between true and false seed matches by means of conservation is less suited in the coding region due to its high general conservation.

4.3.4 Combination of prediction methods

How can the performance of miRNA target site prediction be improved? One way would be to integrate new features that enable significant improvement of precision. An alternative approach that is commonly applied is to combine different methods relying on different assumptions. In contrast to finding new target site features, this approach could be more easily realized. A couple of databases have been published already that integrate target predictions from multiple methods as reviewed by Min et al. [144]. These integration approaches focused on miRNA-

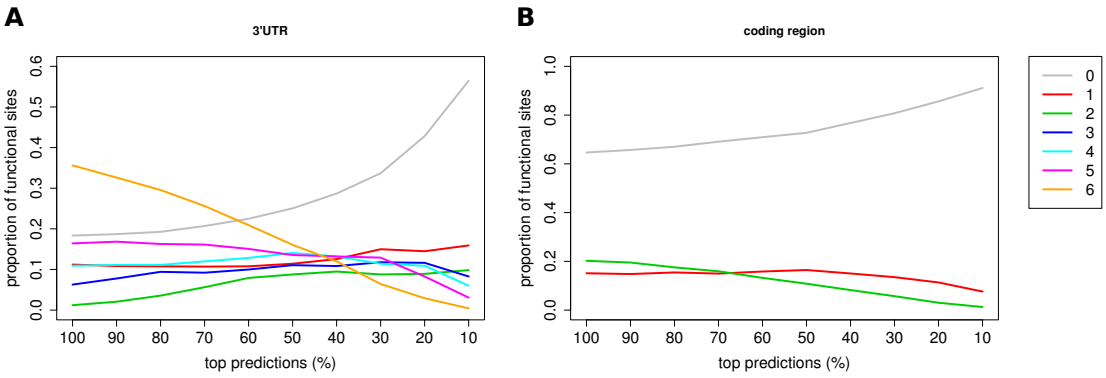


Figure 4.7: INTERSECTION OF PREDICTION RESULT SETS. The occurrence of target sites in different result sets was counted for increasing score cutoffs. Starting with all predictions from all tools, i.e. 100%, the result sets were successively reduced up to the best 10% of predictions according to the tool-specific scores, respectively. For each step the proportions of functional sites covered by 0, 1,..., 6 methods with respect to the 3'UTR (A) and 0, 1, and 2 tools with respect to the coding region (B) have been determined. Note: the ranges of the y-axes of the two plots are different.

target interactions. Whether the combination of various prediction methods on target site level, see figure 4.1, is promising was examined here. Practically, do the results sets of individual tools overlap or are they disjunct? Further, how does the intersection of tools change for increasing score cutoffs? If the prediction methods identified different subsets of functional sites, sensitivity could be largely improved by combining them. The naive PhastCons-based approach was not considered in this analysis.

Figure 4.7(A) shows that the prediction result sets for the 3'UTR diverge, the more poorly rated predictions were removed from them. The initial result sets including all predictions overlapped largely, e.g. 36% of the functional sites were predicted by each of the six methods. But when retaining only the top 30% of predictions of each method, most of the identified functional sites were detected only by one method. Moreover, less than 10% of the functional sites were predicted by three or more tools when considering the top 10% of predictions. Consequently, the higher the quality was of the result sets, the less the overlap was between them suggesting individual approaches or subsets of methods to be complementary to each other. The combination of the top 10% of predictions of all tools achieved a F-measure of 0.19 just as the filtered miRmap set. Even though the precision of

the aggregated set was lower (0.12 vs. 0.17), the sensitivity was more than twice as high as that of miRmap (0.44 vs. 0.20).

With respect to the coding region a similar observation as for the 3'UTR could be made, see figure 4.7(B): The proportion of sites detected by one method exceeded that found by both tools for high cutoffs. Aggregating the top 50% of predictions of both methods yielded the best performing combined result set (F-measure: 0.13). It slightly outperformed EIMMo in terms of the F-measure, but performed worse than Diana-microT. But the sensitivity of the combined set exceeded both single methods.

Concluding, these observations motivate a deeper investigation of this subject. An important question is, which tool combination performs best for the 3'UTR. It is conceivable that a subset of the six tools performs better that the union of all tools that has been evaluated above. For instance, excluding PITA that performed worst of all published approaches, see section 4.3.3, should increase the performance. To explore this all, 2^6 possible combinations should be tested for varying cutoffs. Further, it would be interesting to find out whether there exists a tool combination that outperforms mirSVR.

4.4 Conclusion

This study of target site prediction methods had two goals. The first objective was to clarify which types of seeds, particularly which of the short ones, are considered by a group of representative and frequently used methods. In the second part, these methods were examined how well their scoring functions differentiate between functional and non-functional sites in general.

In the previous chapter, the essential part of a miRNA target site, the seed match, had been described in detail. Here, it was analyzed to what extent these findings are already implemented by target site prediction approaches. Briefly summarized, a substantial proportion of experimentally identified target sites includes short 6mer seed matches. In particular, the 6merα seed type was not considered by previous seed models [7]. This is noteworthy, in particular with respect to coding region, where 6mer sites accounted even for a higher proportion than in the 3'UTR. For this reason, the first part of this evaluation is interesting for those who aim for a comprehensive decoding of the miRNA-target interactome and wonder how this can be achieved with current methods. Against the background of the ceRNA hypothesis, this concern is quite relevant. Salmena et al. motivated such an analysis like this explicitly [108]. The ceRNA hypothesis adds to the primary understanding of a miRNA target site as a cis-regulatory element through which the RNA molecule can be directly regulated a second function, namely, the indirect

regulation of other RNA molecules by sponging miRNAs.

Prediction methods search for 6mer sites but not as comprehensive as for long sites. Most intensively 6merβ sites were searched. Except EIMMo, all approaches contained 6merβ sites in their initial seed match sets. PITA and mirSVR additionally detected 6merγ sites. Diana-microT had almost 50% of the functional 6merα sites in its initial set. But both focusing on specific sequence signals (Diana-microT, EIMMo) and prediction of repression strength (mirSVR, miRmap and TargetScan) resulted in a higher weighting of long seed matches in the prediction of 3'UTR-located target sites. Merely, the 7merα was handled differently by these approaches, see figure 4.9. Consequently, short seeds matches were given the lowest scores and thus these were the first to fall from the result set when the cutoff was increased. The ORF methods showed a similar prioritization of long seed matches, e.g. the ORF variant of Diana microT weighted long seeds even stronger than the 3'UTR variant. Thus, miRNA target site prediction methods are currently not suitable for a full explanation of miRNA-target interactions. Further characteristics describing in particular short sites need to be identified and new approaches have to be developed.

So far in the development of methods the prediction of the direct, cis-regulatory effect of a miRNA target site played the dominating role. This is also reflected in previous evaluations of prediction tools: miRNA-target site interactions were reduced on miRNA-target interactions. Which site or the location of the site causing the interaction was of secondary interest compared to the resulting miRNA-target interaction that could be compared against a reference data set. Another reason for this simplification was perhaps the unavailability of appropriate reference data until recently such as the one used here. However, the position of the target site is required in order to validate a prediction experimentally. Moreover, in the event that a miRNA has multiple potential target sites on a target sequence, it may be important to distinguish between them. For these reasons, an evaluation on target site level as carried out here is of interest.

The analysis of scoring functions used for the prediction in the 3'UTR showed clearly that methods involving several features achieved the highest precision values. mirSVR outperformed any other method for each score cutoff. Surprisingly, miRmap compared poorly with other multiple-feature methods although it was the most recent approach considered here. Its performance even fell below that of Diana-microT and EIMMo for higher sensitivity values. Single-feature methods were only competitive if the scoring function involved signal-to-background calculations. With regard to the coding region, the ORF release of Diana-microT performed best. This was not surprising as Diana-microT was the only method particularly adapted to the coding region.

Finally, the analysis of the complementarity of the methods provided arguments to study the combination of various tools on target site level in detail in order to increase sensitivity of miRNA target site prediction.

4.5 Appendix

	3'UTR		coding region	
	Fun.	Non-fun.	Fun.	Non-fun.
benchmark	1,401	40,982	2,116	57,544
Diana-microT	763	12,239	439	5,527
EIMMo	727	7956	738	10,929
miRmap	974	17,130	-	-
mirSVR	958	15,870	-	-
PITA	1,060	21,602	-	-
TargetScan	729	7,973	-	-

Table 4.3: Benchmark data statistics. This table shows the numbers of functional and non-functional sites per prediction tool.

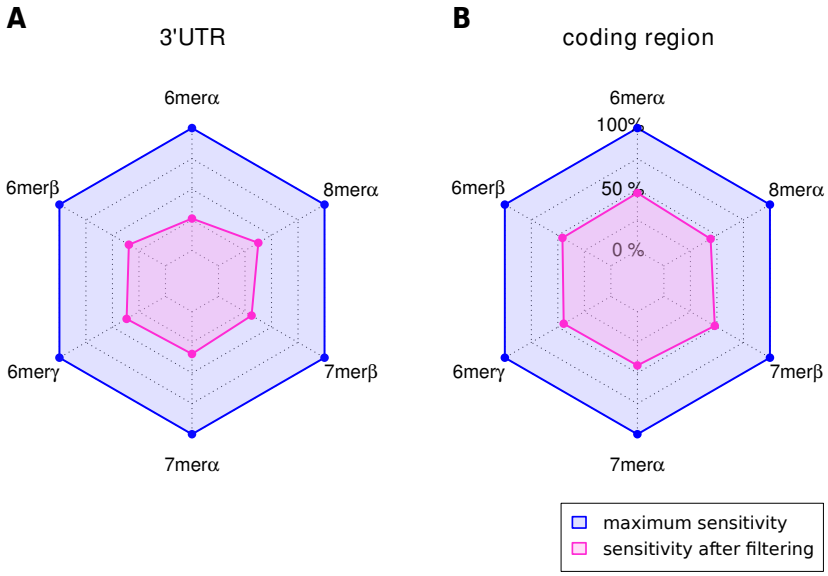


Figure 4.8: Sensitivity of the naive approach per seed type. Captions analogous to figure 4.2.

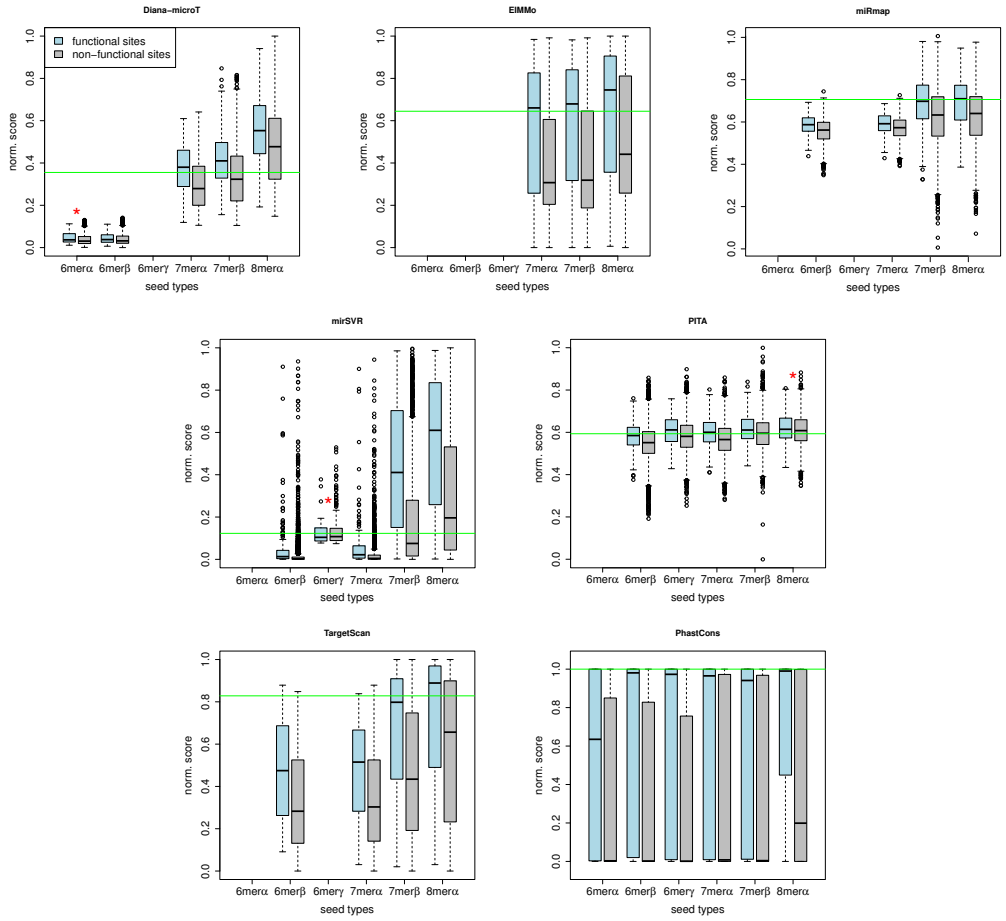


Figure 4.9: Target site score distribution per seed type (3'UTR). Scores were transformed by min-max normalization. Green lines indicate the threshold value t. Red stars highlight score distributions of functional and non-functional sites that do not significantly vary from each other.

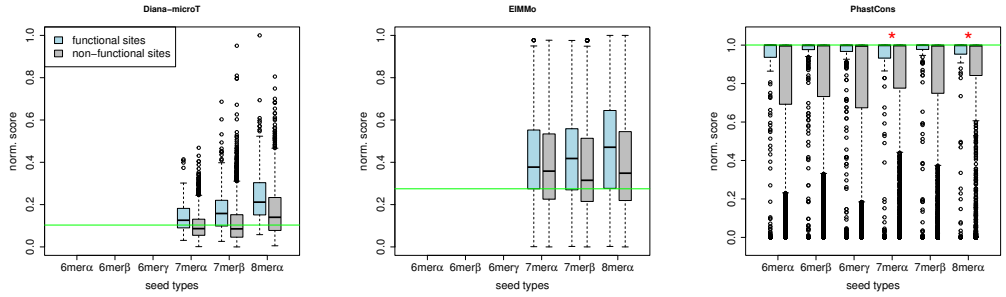


Figure 4.10: Target site score distribution per seed type (coding region). Caption analogous to figure 4.9.

5 Current projects

My current projects deal mainly with the interaction between miRNAs and the coding region. Further, in a collaboration with the IDG at the Helmholtz Zentrum München we are investigating if the Parkinson's disease-related Lrrk2 gene is target of miRNA control. The subsequently presented studies were not far enough progressed to be presented in the result sections of this thesis.

5.1 Frame-variant occurrence of miRNA target sites in the ORF

In both 3'UTR and coding region most target sites start 8 nt downstream of the center of the CCRs, see figure 3.7. But while in the 3'UTR the maximum frequency at position 8 exceeded clearly the proximate frequency values, the differences in ORF-located CCRs were less strong: The numbers of starts at positions 7 and 8 did not differ that much. It was hypothesized that restrictions due to the genetic code appear here. The encoding of the amino acid sequence and the high selective pressure to maintain the sequence limits the scope for the evolution of additional sequence elements. However, the selective pressure varies between the three positions of a codon. According to the genetic code, substituting the third position does not lead to an amino acid exchange in most cases. For this reason, the third positions of codons are weaker conserved compared to codon positions 1 and 2. Possibly, the evolution of target sites has adapted to these ORF-specific conditions.

To investigate this question, the target site start frequencies per position were considered depending on the reading frame, see figure 5.1(A). The frame of a CCR was determined by the codon position of the T/C mutation. This separation led to two observations: (i) T/C positions or respectively CCRs were not uniformly distributed on the three codon positions. CCRs in frame 3 - represented by the red curve - were most frequent. After consulting Markus Hafner (first author of [9]), it can be excluded that this observation was due to an experimental bias. (ii) The preferential start position varied between the frames. For CCRs in frame 1 and 2 the most target sites started 8 nt downstream of the center as in the 3'UTR, see for comparison figure 5.1(B). By contrast if the CCR was in frame 3, most sites started at a distance of 7 nt to the center.

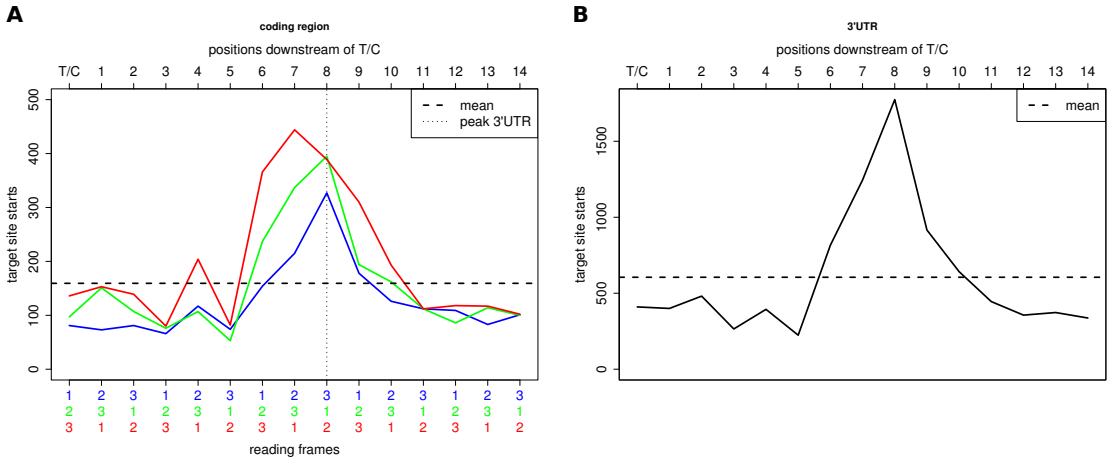


Figure 5.1: Target site starts downstream of the T/C mutation. Target sites starts per position downstream of the central T/C position within CCRs were counted. With regard to the coding region the frequencies were determined separately for each reading frame (A). For comparison the positional frequencies of target sites starts in the 3'UTR are displayed as well (B). Only target sites of the 58 top-expressed miRNAs that made up 95% of the miRNA sequence reads were taken into account for this analysis.

In a first analysis the conservation was compared between these groups of CCRs, subsequently termed CCR1, CCR2 and CCR3 (the number indicates the codon position of the respective T/C mutation). The phyloP algorithm was used to assess the strength of conservation at each site. In contrast to the content-based PhastCons score that takes neighboring bases into account, the phyloP score ignores the conservation of neighbouring bases making it useful for the evaluation of selection at particular sites [145]. Therefore, it is well suited to analyze the codon position-dependent conservation in ORF-located CCRs. Here, the phyloP scores based on a 28-way multiple alignment of vertebrate species were used. The data was downloaded from the UCSC database [117]. The conservation of the first 10 positions downstream of the center of the CCR that are relevant for seed matching were considered (see definition of functional sites in section 3.3.1). Figure 5.2 shows the phyloP score distributions obtained for CCR1, CCR2 and CCR3. Each of the position-specific score distributions was compared with the respective background distribution to test if the conservation at the particular site was significantly stronger or weaker. Considering the background score distributions revealed that most of the third codon positions substitute according to the neutral evolution

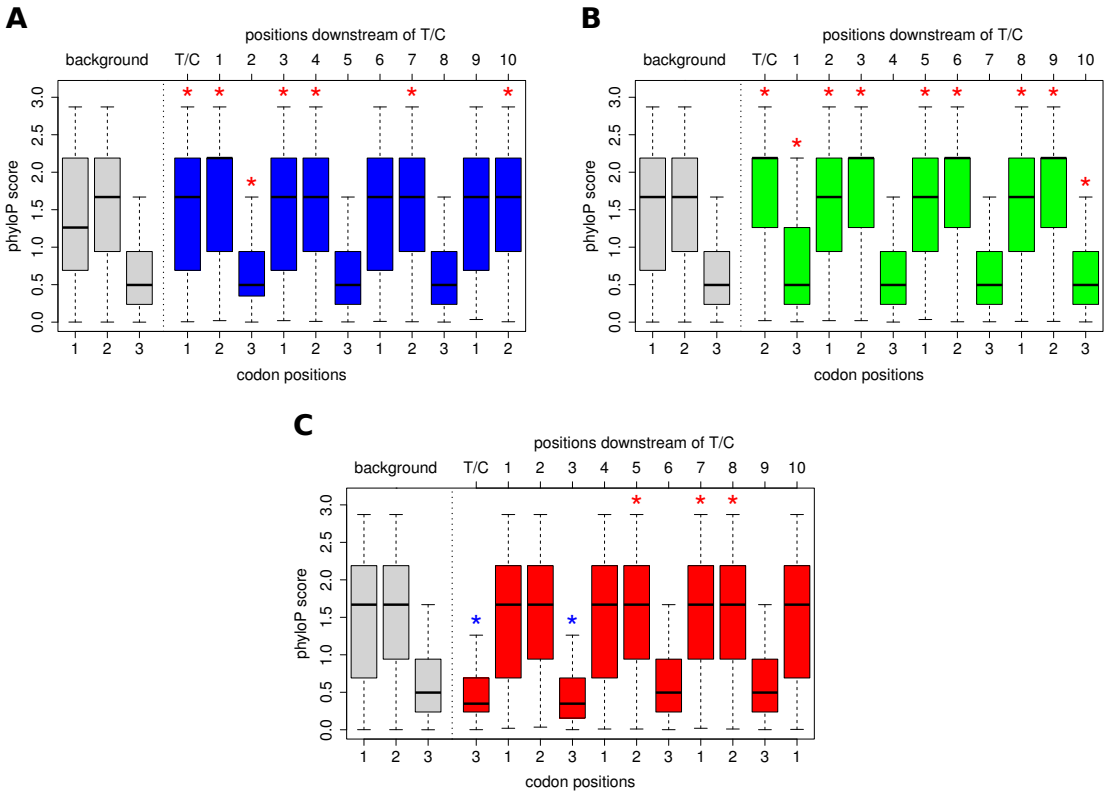


Figure 5.2: Frame-dependent site conservation within CCRs. phyloP measures the conservation at individual alignment columns and ignores the effects of the neighboring columns. The scores represent $-log10$ P-values under a null hypothesis of neutral evolution. Figure (A) displays the phyloP scores of sites downstream of T/C mutations coinciding with codon position 1. Analogously, figures (B) and (C) show the scores for the frames 2 and 3, respectively. The background score distributions are based on all codons included in the subset of coding regions that have minimum one CCR whose central T/C mutation coincides with the respective frame. Stars highlight significant deviation from the respective background codon phyloP score distribution (Wilcoxon test with Bonferroni correction). If the mean was below that of the background distribution, the star was colored blue, otherwise red.

while first and second positions are generally maintained by purifying selection. The selective pressure acting on the second position was strongest. Interestingly, the conservation of individual sites differed between CCR1, CCR2 and CCR3. The

positions T/C-4 in CCR1, 5.2(A), and T/C-3 in CCR2, 5.2(B), turned out to be subject to stronger conservation than the background whereas positions T/C-4 in CCR3, 5.2(C), were similar or even less conserved. Consequently, the increased phyloP scores in both CCR1 and CCR2 regions points to specific conservation of target sites beside the genetic code within the corresponding CCRs. The greatest proportion of CCRs were of type CCR3, 40%, (CCR1: 25%, CCR2: 35%). Possibly, these CCR3 regions are especially appropriate to host species-specific target sites. Further analyses have to be carried out to explain these observations.

5.2 Potential miRNA-mediated regulation of Lrrk2

In a collaboration with Florian Giesert from the IDG we are investigating the post-transcriptional regulation of the *Leucine-rich repeat kinase 2* (*Lrrk2*) gene through miRNAs. At the IDG, *Lrrk2* is explored in mice in context of studying Parkinsons's disease. Parkinson's disease is a neurodegenerative disorder resulting from the death of dopamine-producing neurons [146], [147]. In the analysis of *Lrrk2* expression in adult brain a discrepancy was observed between mRNA and protein level in some regions. Particularly, in the striatum the relative concentration of the LRRK2 protein was less than expected from the mRNA expression [148].

Post-transcriptional regulation by miRNAs was hypothesized to be the cause for this observation. To test this assumption, the 3'UTR of *Lrrk2* was scanned for potential miRNA target sites of all murine miRNAs known at this time (miRBase version 15). The set of seed types displayed in figure 3.8 has been considered. In contrast to the evaluation of prediction methods presented in chapter 4 where the ability to correctly identify the location of target sites has been assessed, here, a prediction on target level has been conducted, see figure 4.1, to get first a set of potential regulators. In total, seed matches of 177 different miRNAs were found on the target sequence. As each candidate miRNA should be tested within a single experiment, only a small subset including very reliable candidates could be checked. To rank the candidates, the seed matches of each miRNA were analyzed. Particularly, the number of seed matches within the *Lrrk2* 3'UTR and the proximity of the seed matches to the stop codon were taken into account. Additionally, the conservation of the sites was determined. Further, the miRNA tissue expression atlas provided by Landgraf et al. [149] was consulted to check whether the candidate miRNAs are expressed in brain. Members of the miR-30 family, including miRNAs miR-30a-e, turned out to be most promising. miR-30b has been chosen for the experiment as it had beside a 8merα seed match close to the stop codon (as all other members of this family as well) the strongest pairing between the 3' region of the miRNA and the target. miR-186 was selected as second candidate for

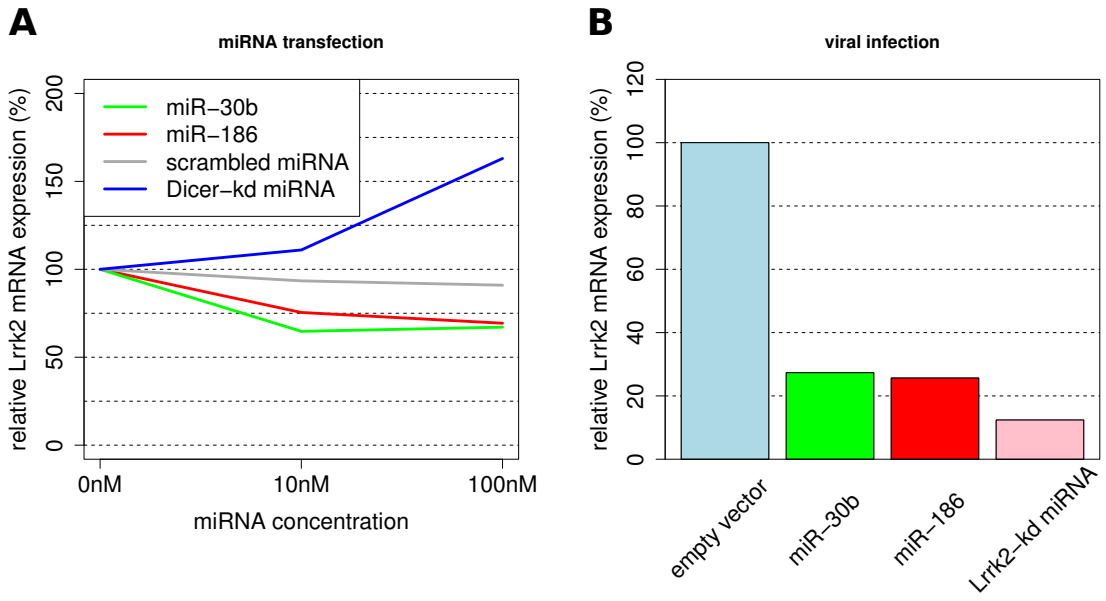


Figure 5.3: Repression of Lrrk2 mRNA expression. (A) miR-186 and miR-30b were transfected into mouse embryonic fibroblasts. Additionally, a miRNA (Dicer-kd, kd: knockdown) specifically downregulating the miRNA processing protein Dicer was added. A scrambled miRNA was used as negative control. (B) An empty viral vector was used as negative control. The pre-miR-186 and pre-miR-30b were integrated into the genome of NIH/3T3 fibroblasts facilitating stable expression. As a positive control an artificial miRNA (Lrrk2-kd) that downregulates *Lrrk2* mRNA was used. The data were kindly provided by Florian Giesert from IDG.

experimental testing. The miR-186 seed matches were weaker conserved than that of miR-30b, but two of them, a 6merβ match and a 7merβ match, were residing in the relevant region close to the stop codon ($< 33\%$ of 3'UTR length). Additional support for miR-186 came from the literature: It is known that miR-186 regulates the expression of the *Foxo1* gene [150]. Further, the LRRK2 protein phosphorylates and thus activates FOXO1 [151]. Possibly, miR-186, *Lrrk2* and *Foxo1* form a regulatory *feed-forward* loop, see [152] for details, with miR-186 negatively regulating *Lrrk2* and *Foxo1* and LRRK2 activating FOXO1.

First, it was tested in mouse embryonic fibroblasts that do endogenously express *Lrrk2* whether interaction between *Lrrk2* and miR-30b or miR-186 could be measured. The impact on the *Lrrk2* mRNA expression level was determined for different concentrations of miRNAs. Figure 5.3(A) shows that both miR-186 and

miR-30b lower the mRNA level by approximately 25% compared to the negative control (scrambled miRNA). Additionally, the Dicer protein that is involved in miRNA processing was knocked down by a miRNA specifically regulating Dicer. Downregulation of Dicer led to an increase of the *Lrrk2* level suggesting that endogenous *Lrrk2* expression is miRNA-dependent. These first results supported the assumption that *Lrrk2* is post-transcriptionally regulated by miRNAs and that miR-186 as well as miR-30b are involved in this regulation. But a significant reduction of *Lrrk2* mRNA level could not be achieved by the transfection experiment. A drawback was that transfected miRNAs were rapidly degraded and thus only short-term effects could be measured. Additionally, there is also the possibility that the large amounts of supplied miRNAs induce artificial stress effects that change the homoeostasis of the cells.

For this reason in a follow-up experiment the miRNAs were delivered by a viral vector into an established fibroblast cell line (NIH/3T3). The vector contained the entire pre-miRNA sequence that was integrated into the genome of the transfected cell. The integrated miRNA was expressed by the transfected cell ensuring an uniform miRNA expression for a longer period. *Lrrk2* mRNA expression was determined seven days after the transfection. Both in miR-30b and miR-186 transfected cells a considerable reduction of the *Lrrk2* mRNA level between 70% and 80% could be observed, see figure 5.3(B). Surprisingly, when the LRRK2 protein level was measured, no reduction of the protein level could be observed. An explanation could be the long half-live of the examined protein as well as compensatory effects on the post-translational level. But since technical reasons can not be fully excluded as well, the experiments are currently repeated in order to stabilize the current findings. In the event that miRNA-Lrrk2 interaction can be experimentally verified in fibroblast cell lines, miRNA regulation will be analyzed in the originally examined brain regions.

6 Conclusion and outlook

The goal of this thesis was to explore two important questions related to computational research of miRNA biology. First, an in-depth analysis of an experimentally determined large-scale miRNA-target interaction map has been carried out to study the forms of miRNA target recognition patterns. Previous studies on that subject primarily pointed to the importance of 7 or 8 nt long sites, i.e. seed matches, on the target that match perfectly to the seed region of the miRNA as reviewed by Bartel [7]. But these studies took mainly conserved seed matches into account. Here, the data basis for the analysis included conserved as well as non-conserved sites. Among the less conserved sites, neither 7mers nor 8mers constituted the majority, but short sites of 6 nt. This class of seed-complementary sites has received little or no attention until previously as they are exceeded by long sites in terms of efficiency and target specificity. But short sites are numerous; within the coding region they occur more than in the 3'UTR. In this work, a possible role for short 6mer hits as miRNA regulatory elements was contemplated: Like other regulatory molecules, miRNAs are not expressed constantly and durably. miRNAs that may not act in a certain stage can be degraded. The hypothesis formulated here suggests that alternatively or in parallel to degradation the regulatory function can be restricted by increasing the concentration of pseudotargets or ceRNAs that contain 6mer seed matches of miRNAs to be deactivated. That is, transcriptome-wide the proportion of target sites including short seed matches rises for this group of miRNAs.

As shown in figures 3.9 and 3.11, interaction with 6mer seed matches has relatively little consequences for the miRNA-bound RNAs. But the miRNA associated to a 6mer site is distracted from pairing with 7mer or 8mer sites that, on the other hand, would induce efficient target repression. The regulation of miRNA function by sponge-like mechanisms as formulated by the ceRNA hypothesis [108] gained greatly in importance by the discovery of circular RNAs. Just recently, these so-called circRNAs were described to contain a multitude of conserved miRNA binding sites [153],[154]. miRNAs bound to circRNAs can not regulate their actual targets and thus are disabled. An interesting future project would be to investigate if and to what extent 6mer sites are modulating miRNA function. First, the observation has to be confirmed with another data set. The next step would

be to analysis expression profiles of miRNAs and mRNAs based on different time points or varying conditions and check if the reservoir of seed matches (long or short sites) correlates with the abundance (high or low) of the matching miRNAs.

In terms of miRNA targeting in the coding region, some new issues were raised. Two studies have shown that ORF-located miRNA target sites support targeting through the 3'UTR [103],[104]. But most of the transcripts with CCRs in the ORF did not include CCRs in the 3'UTR. Consequently, ORF-based targeting occurs also regardless of the 3'UTR. But what is the role of the coding region in miRNA biology? The repression through ORF targets sites is low compared to that of target sites located in the 3'UTR. Indeed, 6mer sites accounted for a larger proportion in the ORF, but this can not explain the diminished efficacy as the target destabilization was less for all seed match lengths in the ORF. So do ORF seed matches mainly carry out ceRNA function? As the coding region has been considered less until now with respect to miRNA-mediated regulation, many new insights can be expected for the future. A point I am especially interested in is to analyze the impact of the genetic code on the occurrence of target sites. Possibly, the genetic code is responsible for the decreased efficacy of ORF target sites. The high selective pressure on the genetic code potentially impairs the formation of efficient target sites. Therefore, an interesting question for the future is, how far is miRNA function in the ORF affected by the genetic code?

From the analysis of miRNA target site predictions algorithms one could conclude that there is still much room for improvement. With respect to the 3'UTR, an acceptable precision (> 0.33) was achieved only at very low sensitivity (< 0.1) by some of the more complex approaches. To be fair, it has to be admitted that apart from Diana-microT none of the methods considers 6merα. But seed types 6merβ and 6merγ that are well known were either not searched or sites of these types were only poorly rated, see figure 4.2. Thus, short seed matches were the first to be removed from the result sets in case restrictive cutoffs were used. This is due to the fact that the methods are geared to predict interactions that can be confirmed with high probability in subsequent experiments. One way to improve the performance was presented: Since the methods considered here produced sets of target sites that did overlap only partially, the sensitivity could be increased by combining the result sets. In principle, however, the known characteristics of functional target sites are not sufficient to comprehensively separate true from false candidates. Biochemical methods for the detection of miRNA target sites are constantly being improved and thus they will allow for a more sensitive identification at high specificity. Potentially, additional as yet unknown features can be derived from these data obtained in the future.

Bibliography

[1] Daniel C Ellwanger, Florian A Büttner, Hans-Werner Mewes, and Volker Stümpflen. The sufficient minimal set of mirna seed types. *Bioinformatics*, 27(10):1346–1350, May 2011.

[2] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[3] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. *Cell*, 75(5):855–862, Dec 1993.

[4] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–858, Oct 2001.

[5] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. *Science*, 294(5543):858–862, Oct 2001.

[6] R. C. Lee and V. Ambros. An extensive class of small rnas in caenorhabditis elegans. *Science*, 294(5543):862–864, Oct 2001.

[7] David P Bartel. Microrna: target recognition and regulatory functions. *Cell*, 136(2):215–233, Jan 2009.

[8] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute hits-clip decodes microrna-mrna interaction maps. *Nature*, 460(7254):479–486, Jul 2009.

[9] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide

identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–141, Apr 2010.

[10] G. Ruvkun, V. Ambros, A. Coulson, R. Waterston, J. Sulston, and H. R. Horvitz. Molecular genetics of the caenorhabditis elegans heterochronic gene lin-14. *Genetics*, 121(3):501–516, Mar 1989.

[11] V. Ambros and H. R. Horvitz. The lin-14 locus of caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. *Genes Dev*, 1(4):398–414, Jun 1987.

[12] V. Ambros. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in c. elegans. *Cell*, 57(1):49–57, Apr 1989.

[13] B. Wightman, T. R. Brglin, J. Gatto, P. Arasu, and G. Ruvkun. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during caenorhabditis elegans development. *Genes Dev*, 5(10):1813–1824, Oct 1991.

[14] Gary Ruvkun, Bruce Wightman, and Ilho Ha. The 20 years it took to recognize the importance of tiny rnas. *Cell*, 116(2 Suppl):S93–6, 2 p following S96, Jan 2004.

[15] Victor Ambros. The evolution of our thinking about micrornas. *Nat Med*, 14(10):1036–1040, Oct 2008.

[16] M. Wickens and K. Takayama. Rna. deviants–or emissaries. *Nature*, 367(6458):17–18, Jan 1994.

[17] E. G. Moss, R. C. Lee, and V. Ambros. The cold shock domain protein lin-28 controls developmental timing in c. elegans and is regulated by the lin-4 rna. *Cell*, 88(5):637–646, Mar 1997.

[18] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–811, Feb 1998.

[19] A. J. Hamilton and D. C. Baulcombe. A species of small antisense rna in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952, Oct 1999.

[20] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, Feb 2000.

[21] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Mller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. *Nature*, 408(6808):86–89, Nov 2000.

[22] F. J. Slack, M. Basson, Z. Liu, V. Ambros, H. R. Horvitz, and G. Ruvkun. The lin-41 rbcc gene acts in the c. elegans heterochronic pathway between the let-7 regulatory rna and the lin-29 transcription factor. *Mol Cell*, 5(4):659–669, Apr 2000.

[23] G. Ruvkun. Molecular biology. glimpses of a tiny rna world. *Science*, 294(5543):797–799, Oct 2001.

[24] T. Ryan Gregory. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet*, 6(9):699–708, Sep 2005.

[25] Ladeana W Hillier, Alan Coulson, John I Murray, Zhirong Bao, John E Sulston, and Robert H Waterston. Genomics in c. elegans: so many genes, such a little worm. *Genome Res*, 15(12):1651–1660, Dec 2005.

[26] Jean Louis Gunet. The mouse genome. *Genome Res*, 15(12):1729–1740, Dec 2005.

[27] Michael Ashburner and Casey M Bergman. Drosophila melanogaster: a case study of a model genomic sequence and its consequences. *Genome Res*, 15(12):1661–1667, Dec 2005.

[28] D. Baltimore. Our genome unveiled. *Nature*, 409(6822):814–816, Feb 2001.

[29] Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, Jul 2003.

[30] Wenyuan Li, Shihua Zhang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–2466, Oct 2012.

[31] Johan H Gibcus and Job Dekker. The context of gene expression regulation. *F1000 Biol Rep*, 4:8, 2012.

[32] Orna Dahan, Hila Gingold, and Yitzhak Pilpel. Regulatory mechanisms and networks couple the different phases of gene expression. *Trends Genet*, 27(8):316–322, Aug 2011.

[33] Juan Mata, Samuel Marguerat, and Jrg Bhler. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci*, 30(9):506–514, Sep 2005.

[34] F. JACOB, D. PERRIN, C. SANCHEZ, and J. MONOD. [operon: a group of genes with the expression coordinated by an operator]. *C R Hebd Seances Acad Sci*, 250:1727–1729, Feb 1960.

[35] John S Mattick. Challenging the dogma: the hidden layer of non-protein-coding rnas in complex organisms. *Bioessays*, 25(10):930–939, Oct 2003.

[36] Vivian L MacKay, Xiaohong Li, Mark R Flory, Eileen Turcott, G. Lynn Law, Kyle A Serikawa, X. L. Xu, Hookeun Lee, David R Goodlett, Ruedi Aebersold, Lue Ping Zhao, and David R Morris. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Mol Cell Proteomics*, 3(5):478–489, May 2004.

[37] Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 13(4):227–232, Apr 2012.

[38] Christine Vogel, Raquel de Sousa Abreu, Daijin Ko, Shu-Yun Le, Bruce A Shapiro, Suzanne C Burns, Devraj Sandhu, Daniel R Boutz, Edward M Marcotte, and Luiz O Penalva. Sequence signatures and mrna concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol*, 6:400, Aug 2010.

[39] Bjrn Schwanhusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011.

[40] Rickard Sandberg, Joel R Neilson, Arup Sarma, Phillip A Sharp, and Christopher B Burge. Proliferating cells express mrnas with shortened 3' un-

translated regions and fewer microrna target sites. *Science*, 320(5883):1643–1647, Jun 2008.

[41] Christine Mayr and David P Bartel. Widespread shortening of 3'utrs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, Aug 2009.

[42] Bjrn Schwanhusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Corrigendum: Global quantification of mammalian gene expression control. *Nature*, Feb 2013.

[43] Eric Huntzinger and Elisa Izaurralde. Gene silencing by micrornas: contributions of translational repression and mrna decay. *Nat Rev Genet*, 12(2):99–110, Feb 2011.

[44] Matthias Selbach, Bjrn Schwanhusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by micrornas. *Nature*, 455(7209):58–63, Sep 2008.

[45] Daehyun Baek, Judit Villn, Chanseok Shin, Fernando D. Camargo, Steven P. Gygi, and David P. Bartel. The impact of micrornas on protein output. *Nature*, 455(7209):64–71, Sep 2008.

[46] Shankar Mukherji, Margaret S. Ebert, Grace X Y. Zheng, John S. Tsang, Phillip A. Sharp, and Alexander van Oudenaarden. Microrna can generate thresholds in target gene expression. *Nat Genet*, 43(9):854–859, Sep 2011.

[47] Aaron Arvey, Erik Larsson, Chris Sander, Christina S Leslie, and Debora S Marks. Target mrna abundance dilutes microrna and sirna activity. *Mol Syst Biol*, 6:363, Apr 2010.

[48] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, Jan 2004.

[49] Alex Mas Monteys, Ryan M Spengler, Ji Wan, Luis Tecedor, Kimberly A Lennox, Yi Xing, and Beverly L Davidson. Structure and activity of putative intronic mirna promoters. *RNA*, 16(3):495–505, Mar 2010.

[50] Jakub O Westholm and Eric C Lai. Mirtrons: microrna biogenesis via splicing. *Biochimie*, 93(11):1897–1904, Nov 2011.

[51] Richard W Carthew and Erik J Sontheimer. Origins and mechanisms of mirnas and sirnas. *Cell*, 136(4):642–655, Feb 2009.

[52] Corine T Neilsen, Gregory J Goodall, and Cameron P Bracken. Isomirs–the overlooked repertoire in the dynamic micrornaome. *Trends Genet*, 28(11):544–549, Nov 2012.

[53] V. Narry Kim. Microrna biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*, 6(5):376–385, May 2005.

[54] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microrna biogenesis, function and decay. *Nat Rev Genet*, 11(9):597–610, Sep 2010.

[55] Stefan Regger and Helge Grohans. Microrna turnover: when, how, and why. *Trends Biochem Sci*, 37(10):436–446, Oct 2012.

[56] Jacek Krol, Volker Busskamp, Ilona Markiewicz, Michael B Stadler, Sebastian Ribi, Jens Richter, Jens Duebel, Silvia Bicker, Hans Jrg Fehling, Dirk Schbeler, Thomas G Oertner, Gerhard Schratt, Miriam Bibel, Botond Roska, and Witold Filipowicz. Characterizing light-regulated retinal microrna reveals rapid turnover as a common property of neuronal microrna. *Cell*, 141(4):618–631, May 2010.

[57] Zoya S Kai and Amy E Pasquinelli. Microrna assassins: factors that regulate the disappearance of mirnas. *Nat Struct Mol Biol*, 17(1):5–10, Jan 2010.

[58] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res*, 37(Database issue):D98–104, Jan 2009.

[59] Jongpil Kim, Keiichi Inoue, Jennifer Ishii, William B. Vanti, Sergey V. Voronov, Elizabeth Murchison, Gregory Hannon, and Asa Abeliovich. A microrna feedback circuit in midbrain dopamine neurons. *Science*, 317(5842):1220–1224, Aug 2007.

[60] Richard I. Gregory, Kai-Ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar. The microprocessor complex mediates the genesis of microrna. *Nature*, 432(7014):235–240, Nov 2004.

[61] Nila Roy Choudhury and Gracjan Michlewski. Terminal loop-mediated control of microrna biogenesis. *Biochem Soc Trans*, 40(4):789–793, Aug 2012.

[62] Joshua J Forman, Aster Legesse-Miller, and Hilary A Coller. A search for conserved sequences in coding regions reveals that the let-7 microrna targets dicer within its coding sequence. *Proc Natl Acad Sci U S A*, 105(39):14879–14884, Sep 2008.

[63] Sarath Chandra Janga. From specific to global analysis of posttranscriptional regulation in eukaryotes: posttranscriptional regulatory networks. *Brief Funct Genomics*, 11(6):505–521, Nov 2012.

[64] Nicole-Claudia Meisner and Witold Filipowicz. Properties of the regulatory rna-binding protein hur and its role in controlling mirna repression. *Adv Exp Med Biol*, 700:106–123, 2010.

[65] Amy E Pasquinelli. Micrornas and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet*, 13(4):271–282, Apr 2012.

[66] Martijn Kedde, Marieke van Kouwenhove, Wilbert Zwart, Joachim A F. Oude Vrielink, Ran Elkon, and Reuven Agami. A pumilio-induced rna structure switch in p27-3' utr controls mir-221 and mir-222 accessibility. *Nat Cell Biol*, 12(10):1014–1020, Oct 2010.

[67] Hun-Way Hwang, Erik A. Wentzel, and Joshua T. Mendell. A hexanucleotide element directs microrna nuclear import. *Science*, 315(5808):97–100, Jan 2007.

[68] Alessia Baccarini, Hemangini Chauhan, Thomas J Gardner, Anitha D Jayaprakash, Ravi Sachidanandam, and Brian D Brown. Kinetic analysis reveals the fate of a microrna following target regulation in mammalian cells. *Curr Biol*, 21(5):369–376, Mar 2011.

[69] Saibal Chatterjee, Monika Fasler, Ingo Bssing, and Helge Grosshans. Target-mediated protection of endogenous micrornas in c. elegans. *Dev Cell*, 20(3):388–396, Mar 2011.

[70] Richard J Jackson, Christopher U T Hellen, and Tatyana V Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol*, 11(2):113–127, Feb 2010.

[71] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mrna translation and stability by microrna. *Annu Rev Biochem*, 79:351–379, 2010.

[72] Shobha Vasudevan, Yingchun Tong, and Joan A Steitz. Switching from repression to activation: microrna can up-regulate translation. *Science*, 318(5858):1931–1934, Dec 2007.

[73] Shobha Vasudevan. Posttranscriptional upregulation by microrna. *Wiley Interdiscip Rev RNA*, 3(3):311–330, 2012.

[74] P. H. Olsen and V. Ambros. The lin-4 regulatory rna controls developmental timing in caenorhabditis elegans by blocking lin-14 protein synthesis after the initiation of translation. *Dev Biol*, 216(2):671–680, Dec 1999.

[75] Shveta Bagga, John Bracht, Shaun Hunter, Katlin Massirer, Janette Holtz, Rachel Eachus, and Amy E Pasquinelli. Regulation by let-7 and lin-4 mirnas results in target mrna degradation. *Cell*, 122(4):553–563, Aug 2005.

[76] Sergej Djuranovic, Ali Nahvi, and Rachel Green. A parsimonious model for gene regulation by mirnas. *Science*, 331(6017):550–553, Feb 2011.

[77] Marc R Fabian and Nahum Sonenberg. The mechanics of mirna-mediated gene silencing: a look under the hood of mirisc. *Nat Struct Mol Biol*, 19(6):586–593, Jun 2012.

[78] Suvendra N. Bhattacharyya, Regula Habermacher, Ursula Martine, Ellen I. Closs, and Witold Filipowicz. Relief of microrna-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–1124, Jun 2006.

[79] Wenqian Hu and Jeff Coller. What comes first: translational repression or mrna degradation? the deepening mystery of microrna function. *Cell Res*, 22(9):1322–1324, Sep 2012.

[80] Sergej Djuranovic, Ali Nahvi, and Rachel Green. mirna-mediated gene silencing by translational repression followed by mrna deadenylation and decay. *Science*, 336(6078):237–240, Apr 2012.

[81] Ariel A. Bazzini, Miler T. Lee, and Antonio J. Giraldez. Ribosome profiling shows that mir-430 reduces translation before causing mrna decay in zebrafish. *Science*, 336(6078):233–237, Apr 2012.

[82] Huili Guo, Nicholas T Ingolia, Jonathan S Weissman, and David P Bartel. Mammalian micrornas predominantly act to decrease target mrna levels. *Nature*, 466(7308):835–840, Aug 2010.

[83] Eric C Lai. Micro rnas are complementary to 3' utr sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–364, Apr 2002.

[84] Lee P Lim, Nelson C Lau, Earl G Weinstein, Aliaa Abdelhakim, Soraya Yekta, Matthew W Rhoades, Christopher B Burge, and David P Bartel. The micrornas of caenorhabditis elegans. *Genes Dev*, 17(8):991–1008, Apr 2003.

[85] Benjamin P Lewis, I hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, Dec 2003.

[86] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[87] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mrnas are conserved targets of microrna. *Genome Res*, 19(1):92–105, Jan 2009.

[88] Monica C Vella, Eun-Young Choi, Shin-Yi Lin, Kristy Reinert, and Frank J Slack. The c. elegans microrna let-7 binds to imperfect let-7 complementary sites from the lin-41 3'utr. *Genes Dev*, 18(2):132–137, Jan 2004.

[89] Chanseok Shin, Jin-Wu Nam, Kyle Kai-How Farh, H. Rosaria Chiang, Alena Shkumatava, and David P Bartel. Expanding the microrna targeting code: functional sites with centered pairing. *Mol Cell*, 38(6):789–802, Jun 2010.

[90] Naama Elefant, Yael Altuvia, and Hanah Margalit. A wide repertoire of mirna binding sites: prediction and functional implications. *Bioinformatics*, 27(22):3093–3101, Nov 2011.

[91] Dominic Didiano and Oliver Hobert. Perfect seed pairing is not a generally reliable predictor for mirna-target interactions. *Nat Struct Mol Biol*, 13(9):849–851, Sep 2006.

[92] Sung Wook Chi, Gregory J Hannon, and Robert B Darnell. An alternative mode of microrna target recognition. *Nat Struct Mol Biol*, 19(3):321–327, Mar 2012.

[93] Kyle Kai-How Farh, Andrew Grimson, Calvin Jan, Benjamin P Lewis, Wendy K Johnston, Lee P Lim, Christopher B Burge, and David P Bartel. The widespread impact of mammalian micrornas on mrna repression and evolution. *Science*, 310(5755):1817–1821, Dec 2005.

[94] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. Microrna targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, Jul 2007.

[95] J. Robin Lytle, Therese A Yario, and Joan A Steitz. Target mrnas are repressed as efficiently by microrna-binding sites in the 5' utr as in the 3' utr. *Proc Natl Acad Sci U S A*, 104(23):9667–9672, Jun 2007.

[96] Matthew W Jones-Rhoades and David P Bartel. Computational identification of plant microrna and their targets, including a stress-induced mirna. *Mol Cell*, 14(6):787–799, Jun 2004.

[97] Isidore Rigoutsos. New tricks for animal microrna: targeting of amino acid coding regions at conserved and nonconserved sites. *Cancer Res*, 69(8):3245–3248, Apr 2009.

[98] G. Pesole, G. Grillo, A. Larizza, and S. Liuni. The untranslated regions of eukaryotic mrnas: structure, function, evolution and bioinformatic tools for their analysis. *Brief Bioinform*, 1(3):236–249, Sep 2000.

[99] Shuo Gu, Lan Jin, Feijie Zhang, Peter Sarnow, and Mark A Kay. Biological basis for restriction of microrna targets to the 3' untranslated region in mammalian mrnas. *Nat Struct Mol Biol*, 16(2):144–150, Feb 2009.

[100] Shalev Itzkovitz and Uri Alon. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res*, 17(4):405–412, Apr 2007.

[101] Michael Schnall-Levin, Yong Zhao, Norbert Perrimon, and Bonnie Berger. Conserved microrna targeting in drosophila is as widespread in coding regions as in 3'utrs. *Proc Natl Acad Sci U S A*, 107(36):15751–15756, Sep 2010.

[102] Pl Saetrom, Bret S E Heale, Ola Snve, Lars Aagaard, Jessica Alluin, and John J Rossi. Distance constraints between microrna target sites dictate efficacy and cooperativity. *Nucleic Acids Res*, 35(7):2333–2342, 2007.

[103] Zhuo Fang and Nikolaus Rajewsky. The impact of mirna target sites in coding sequences and in 3'utrs. *PLoS One*, 6(3):e18067, 2011.

[104] Michael Schnall-Levin, Olivia S Rissland, Wendy K Johnston, Norbert Perrimon, David P Bartel, and Bonnie Berger. Unusually effective microrna targeting within repeat-rich coding regions of mammalian mrnas. *Genome Res*, 21(9):1395–1403, Sep 2011.

[105] Martin Reczko, Manolis Maragkakis, Panagiotis Alexiou, Ivo Grosse, and Artemis G Hatzigeorgiou. Functional microrna targets in protein coding sequences. *Bioinformatics*, 28(6):771–776, Mar 2012.

[106] Ray M. Marn and Ji Vanek. Optimal use of conservation and accessibility filters in microrna target prediction. *PLoS ONE*, 7(2):e32208, 02 2012.

[107] Jean Hausser, Afzal Pasha Syed, Biter Bilen, and Mihaela Zavolan. Analysis of cds-located mirna target sites suggests that they can effectively inhibit translation. *Genome Res*, Jan 2013.

[108] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A cerna hypothesis: the rosetta stone of a hidden rna language? *Cell*, 146(3):353–358, Aug 2011.

[109] Ignacio Rubio-Somoza, Detlef Weigel, Jos-Manuel Franco-Zorilla, Juan Antonio Garca, and Javier Paz-Ares. cernas: mirna target mimic mimics. *Cell*, 147(7):1431–1432, Dec 2011.

[110] Herv Seitz. Redefining microrna targets. *Curr Biol*, 19(10):870–873, May 2009.

[111] Jos Manuel Franco-Zorrilla, Adrin Valli, Marco Todesco, Isabel Mateos, Mara Isabel Puga, Ignacio Rubio-Somoza, Antonio Leyva, Detlef Weigel, Juan Antonio Garca, and Javier Paz-Ares. Target mimicry provides a new mechanism for regulation of microrna activity. *Nat Genet*, 39(8):1033–1037, Aug 2007.

[112] Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature*, 465(7301):1033–1038, Jun 2010.

[113] Marshall Thomas, Judy Lieberman, and Ashish Lal. Desperately seeking microrna targets. *Nat Struct Mol Biol*, 17(10):1169–1174, Oct 2010.

[114] George Easow, Aurelio A Teleman, and Stephen M Cohen. Isolation of microrna targets by mirnp immunopurification. *RNA*, 13(8):1198–1204, Aug 2007.

[115] Jeffry L Corden. Shining a new light on rna-protein interactions. *Chem Biol*, 17(4):316–318, Apr 2010.

[116] Kim D Pruitt, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Res*, 40(Database issue):D130–D135, Jan 2012.

[117] Timothy R Dreszer, Donna Karolchik, Ann S Zweig, Angie S Hinrichs, Brian J Raney, Robert M Kuhn, Laurence R Meyer, Mathew Wong, Cricket A Sloan, Kate R Rosenbloom, Greg Roe, Brooke Rhead, Andy Pohl, Venkat S Malladi, Chin H Li, Katrina Learned, Vanessa Kirkup, Fan Hsu, Rachel A Harte, Luvina Guruvadoo, Mary Goldman, Belinda M Giardine, Pauline A Fujita, Mark Diekhans, Melissa S Cline, Hiram Clawson, Galt P Barber, David Haussler, and W. James Kent. The ucsc genome browser database: extensions and updates 2011. *Nucleic Acids Res*, 40(Database issue):D918–D923, Jan 2012.

[118] Sam Griffiths-Jones. mirbase: microrna sequences and annotation. *Curr Protoc Bioinformatics*, Chapter 12:Unit 12.9.1–Unit 12.910, Mar 2010.

[119] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, Aug 2005.

[120] Doron Betel, Manda Wilson, Aaron Gabow, Debora S Marks, and Chris Sander. The microrna.org resource: targets and expression. *Nucleic Acids Res*, 36(Database issue):D149–D153, Jan 2008.

[121] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[122] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[123] Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, Jun 2006.

[124] Hui Chen and Mathieu Blanchette. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol*, 7 Suppl 1:S9, 2007.

[125] Florian A Büttner, Daniel Ellwanger, Hans-Werner Mewes, and Volker Stümpflen. Large scale analysis reveals novel insights in the characteristics of mirna targeting. In Grote A Schomburg D, editor, *Short Papers*. German Conference on Bioinformatics, 2010.

[126] Ray M Marn, Miroslav Sulc, and Jir Vancek. Searching the coding region for microrna targets. *RNA*, Feb 2013.

[127] Praveen Sethupathy, Molly Megraw, and Artemis G Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microrna targets. *Nat Methods*, 3(11):881–886, Nov 2006.

[128] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L Papadopoulos, Martin Reczko, and Artemis G Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microrna target identification. *Bioinformatics*, 25(23):3049–3055, Dec 2009.

[129] Dong Yue, Hui Liu, and Yufei Huang. Survey of computational algorithms for microrna target prediction. *Curr Genomics*, 10(7):478–492, Nov 2009.

[130] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome Biol*, 11(8):R90, 2010.

[131] Charles E Vejnar and Evgeny M Zdobnov. mirmap: Comprehensive prediction of microrna target repression strength. *Nucleic Acids Res*, 40(22):11673–11683, Dec 2012.

[132] Manolis Maragkakis, Panagiotis Alexiou, Giorgio L Papadopoulos, Martin Reczko, Theodore Dalamagas, George Giannopoulos, George Goumas, Evangelos Koukis, Kornilios Kourtis, Victor A Simossis, Praveen Sethupathy, Thanasis Vergoulis, Nectarios Koziris, Timos Sellis, Panagiotis Tsanakas, and Artemis G Hatzigeorgiou. Accurate microrna target prediction correlates with protein repression levels. *BMC Bioinformatics*, 10:295, 2009.

[133] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, 2007.

[134] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nat Genet*, 39(10):1278–1284, Oct 2007.

[135] David M Garcia, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other micrornas. *Nat Struct Mol Biol*, 18(10):1139–1146, Oct 2011.

[136] Molly Hammell. Computational methods to identify mirna targets. *Semin Cell Dev Biol*, 21(7):738–744, Sep 2010.

[137] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. Microrna targets in drosophila. *Genome Biol*, 5(1):R1, 2003.

[138] Alexander Stark, Julius Brennecke, Robert B Russell, and Stephen M Cohen. Identification of drosophila microrna targets. *PLoS Biol*, 1(3):E60, Dec 2003.

[139] Molly Hammell, Dang Long, Liang Zhang, Andrew Lee, C. Steven Carmack, Min Han, Ye Ding, and Victor Ambros. mirwip: microrna target prediction based on microrna-containing ribonucleoprotein-enriched transcripts. *Nat Methods*, 5(9):813–819, Sep 2008.

[140] Kevin C Miranda, Tien Huynh, Yvonne Tay, Yen-Sin Ang, Wai-Leong Tam, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217, Sep 2006.

[141] Ray M Marn and Jir Vancek. Efficient use of accessibility in microrna target prediction. *Nucleic Acids Res*, 39(1):19–29, Jan 2011.

[142] William H Majoros and Uwe Ohler. Spatial preferences of microrna targets in 3' untranslated regions. *BMC Genomics*, 8:152, 2007.

[143] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas K Khri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Monika Komorowska, Gautier Koscielny, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Matthieu Muffato, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Harpreet Singh Riat, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Y. Amy Tang, Kieron Taylor, Stephen Trevanion, Jana Vandrovcova, Simon White, Mark Wilson, Steven P Wilder, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xos M Fernndez-Suarez, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, Giulietta Spudich, Jan Vogel, Andy Yates, Amonida Zadissa, and Stephen M J Searle. Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–D90, Jan 2012.

[144] Hyeyoung Min and Sungroh Yoon. Got target? computational methods for microrna target prediction and their extension. *Exp Mol Med*, 42(4):233–244, Apr 2010.

[145] Adam Siepel, Katherine S. Pollard, and David Haussler. New methods for detecting lineage-specific selection. In *Proceedings of the 10th annual international conference on Research in Computational Molecular Biology*, RECOMB'06, pages 190–205, Berlin, Heidelberg, 2006. Springer-Verlag.

[146] Alexander Zimprich, Bertram Mller-Myhsok, Matthew Farrer, Petra Leitner, Manu Sharma, Mary Hulihan, Paul Lockhart, Audrey Strongosky, Jennifer Kachergus, Donald B Calne, Jon Stoessl, Ryan J Uitti, Ronald F Pfeiffer, Claudia Trenkwalder, Nikolaus Homann, Erwin Ott, Karoline Wenzel, Friedrich Asmus, John Hardy, Zbigniew Wszolek, and Thomas Gasser. The

park8 locus in autosomal dominant parkinsonism: confirmation of linkage and further delineation of the disease-containing interval. *Am J Hum Genet*, 74(1):11–19, Jan 2004.

[147] Coro Paisn-Ruz, Shushant Jain, E. Whitney Evans, William P Gilks, Javier Simn, Marcel van der Brug, Adolfo Lpez de Munain, Silvia Aparicio, Angel Martnez Gil, Naheed Khan, Janel Johnson, Javier Ruiz Martinez, David Nicholl, Itxaso Marti Carrera, Amets Sanz Pena, Rohan de Silva, Andrew Lees, Jos Flix Mart-Mass, Jordi Prez-Tur, Nick W Wood, and Andrew B Singleton. Cloning of the gene containing mutations that cause park8-linked parkinson's disease. *Neuron*, 44(4):595–600, Nov 2004.

[148] Florian Giesert, Andreas Hofmann, Alexander Brger, Julia Zerle, Karina Kloos, Ulrich Hafen, Luise Ernst, Jingzhong Zhang, Daniela Maria Vogt-Weisenhorn, and Wolfgang Wurst. Expression analysis of lrrk1, lrrk2 and lrrk2 splice variants in mice. *PLoS One*, 8(5):e63778, 2013.

[149] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sbastien Pfeffer, Amanda Rice, Alice O Kamphorst, Markus Landthaler, Carolina Lin, Nicholas D Socci, Leandro Hermida, Valerio Fulci, Sabina Chiaretti, Robin Fo, Julia Schliwka, Uta Fuchs, Astrid Novosel, Roman-Ulrich Mller, Bernhard Schermer, Ute Bissels, Jason Inman, Quang Phan, Minchen Chien, David B Weir, Ruchi Choksi, Gabriella De Vita, Daniela Frezzetti, Hans-Ingo Trompeter, Veit Hornung, Grace Teng, Gunther Hartmann, Miklos Palkovits, Roberto Di Lauro, Peter Wernet, Giuseppe Macino, Charles E Rogler, James W Nagle, Jingyue Ju, F. Nina Papavasiliou, Thomas Benzing, Peter Lichter, Wayne Tam, Michael J Brownstein, Andreas Bosio, Arndt Borkhardt, James J Russo, Chris Sander, Mihaela Zavolan, and Thomas Tuschl. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, Jun 2007.

[150] Stephen S Myatt, Jun Wang, Lara J Monteiro, Mark Christian, Ka-Kei Ho, Luca Fusi, Roberto E Dina, Jan J Brosens, Sadaf Ghaem-Maghami, and Eric W-F Lam. Definition of micrornas that repress expression of the tumor suppressor gene foxo1 in endometrial cancer. *Cancer Res*, 70(1):367–377, Jan 2010.

[151] Tomoko Kanao, Katerina Venderova, David S Park, Terry Unterman, Bingwei Lu, and Yuzuru Imai. Activation of foxo by lrrk2 induces expression of proapoptotic proteins and alters survival of postmitotic dopaminergic neuron in drosophila. *Hum Mol Genet*, 19(19):3747–3758, Oct 2010.

[152] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A*, 100(21):11980–11985, Oct 2003.

[153] Thomas B. Hansen, Trine I. Jensen, Bettina H. Clausen, Jesper B. Bramsen, Bente Finsen, Christian K. Damgaard, and Jrgen Kjems. Natural rna circles function as efficient microrna sponges. *Nature*, 495(7441):384–388, Mar 2013.

[154] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D. Mackowiak, Lea H. Gregersen, Mathias Munschauer, Alexander Loewer, Ulrike Ziebold, Markus Landthaler, Christine Kocks, Ferdinand le Noble, and Nikolaus Rajewsky. Circular rnas are a large class of animal rnas with regulatory potency. *Nature*, 495(7441):333–338, Mar 2013.