

Master Thesis
Bioinformatics and Systems Biology
Albert-Ludwigs-University Freiburg

Improving microRNA target prediction in humans using a highly descriptive graph-based machine-learning model

Michael Uhl ¹

February 13, 2014

Supervisors
Prof. Dr. Rolf Backofen ²
and PD Dr. Bjoern Voss ³

¹E-mail: uhlm@informatik.uni-freiburg.de

²Chair of Bioinformatics, Albert-Ludwigs-University Freiburg

³Institute for Biology III, Albert-Ludwigs-University Freiburg

Acknowledgements

No work can be accomplished without the help of others. I would like to take this time and thank the people who made this one possible:

Prof. Dr. Rolf Backofen and Sita Lange for giving me the opportunity to work on this interesting subject.

Sita Lange and Dr. Fabrizio Costa for their great supervision and patient answering of my interrogating questions.

PD Dr. Bjoern Voss for consenting to be the second supervisor.

Daniel Maticzka and all the other group members for their help and contributions.

My family for their continuing support throughout the years. Thank you for sending me this huge birthday survival package. It saved my life.

My girlfriend Teresa. You hold a special place in my heart ... I guess now it's my turn to cook.

Contents

Acknowledgements	iii
Abstract	vii
1. Introduction	1
1.1. microRNA biology	2
1.1.1. Biogenesis and processing	2
1.1.2. Mode of action	3
1.1.3. Refining the biological model	4
1.2. Methods for microRNA target site identification	5
1.2.1. Computational methods	5
1.2.2. Target-specific methods	5
1.2.3. High-throughput methods	6
1.3. Motivation and objectives	8
1.4. Structure	9
2. Principles of microRNA targeting	10
2.1. Sequence complementarity	10
2.2. Thermodynamic stability	11
2.3. Target site accessibility	12
2.4. Evolutionary conservation	12
2.5. Additional principles	13
3. The graph kernel model	15
3.1. Abstract shapes structure representation	15
3.2. gSpan graph encoding	16
3.3. Graph kernel feature decomposition	18

3.4.	Model training and evaluation	19
3.5.	Performance measures	20
4.	Methods	22
4.1.	Data pre-processing and analysis	22
4.1.1.	mRNA and microRNA sequences	23
4.1.2.	AGO1-4 PAR-CLIP dataset	23
4.1.3.	AGO2 CLIP and PAR-CLIP datasets	24
4.1.4.	AGO1 CLASH dataset	24
4.1.5.	RNA-binding protein mRNA occupancy dataset	25
4.2.	Seed scanning and hybrid prediction	26
4.2.1.	Incorporated seed types	26
4.2.2.	Regular expression seed scanning	27
4.2.3.	IntaRNA hybrid prediction	27
4.3.	Graph generation and extensions	28
4.3.1.	gSpan graph generation	28
4.3.2.	miRNA graph extensions	28
4.4.	Construction of test and training datasets	30
4.4.1.	Positive interactions	30
4.4.2.	Negative interactions	31
4.5.	The computational pipeline	32
5.	Results and Discussion	35
5.1.	Data analysis	35
5.1.1.	mRNA occupancy mapping	36
5.1.2.	Importance of target and miRNA abundance	39
5.1.3.	Further dataset characteristics	41
5.2.	Model selection	44
5.2.1.	Interaction sections	45
5.2.2.	Viewpoints extension	46
5.2.3.	Graph kernel parameter optimization	47
5.3.	Model evaluation	50
5.3.1.	CLIP dataset performances	50
5.3.2.	Assessing generalization ability	51
5.3.3.	Testing the trained models	52
6.	Conclusion and Outlook	55
A.	Computational Details	57
A.1.	Data pre-processing	57
A.1.1.	Sequence feature extraction	57
A.1.2.	Target sequence mapping	58
A.1.3.	BED format operations	59
A.1.4.	Human genome assembly conversion	59

Contents

A.2. Training data generation and utilization	60
A.2.1. Regular expression seed scanning	60
A.2.2. IntaRNA hybrid prediction	61
A.2.3. FASTA to gSpan conversion	62
A.2.4. EDeN feature extraction and model training	62
A.3. Computational pipeline description	63
B. Abbreviations	66

Abstract

Computational prediction of animal microRNA target sites imposes a tough challenge on research, since complementarity of functional microRNA-target interactions is usually small, which inevitably leads to a high number of false positive predictions. Prediction programs try to cope with this dilemma by applying additional filtering, but still their current performances are far from optimal.

In the course of this thesis, a novel graph-based machine learning model was extended in order to be utilized for microRNA target prediction. Recently published high-throughput datasets of microRNA-target interactions have been compiled to train and test the generated models. Furthermore, the datasets have been analysed in order to study microRNA related characteristics.

The principle idea behind the graph-based approach is to encode microRNA-target interactions as graphs, which can be efficiently evaluated using a graph-kernel method in combination with a machine learning model. Moreover, additional features of microRNA interactions can be encoded into the graphs, and their relative contributions on prediction performance can be evaluated.

Regarding the obtained results, model training resulted in good predictive performances, while model testing on an independent dataset still has room for improvement. Moreover, analysis of the datasets revealed some interesting insights which should help to improve future prediction studies, especially when working with the analysed datasets.

CHAPTER 1

Introduction

Looking back, as we witness the 20th anniversary of the initial discovery of microRNAs in 1993, research has come a long way in exploring microRNA functionality throughout animals and plants. Over the years, the number of annotated microRNAs has grown from hundreds to tens of thousands, and so has the number of microRNA-related publications. It is now widely accepted that microRNAs regulate gene expression via binding to designated sites on target mRNAs, predominantly resulting in translational repression, target RNA cleavage or decay. Likewise, microRNAs have been linked to various diseases, stimulating research efforts to better understand the fundamental principles of microRNA targeting.

During the past five years, high-throughput methods have entered the field of microRNA research, supplying scientists with vast collections of transcriptome-wide microRNA target sites for selected cell types. These sets undeniably provide valuable new insights into the mechanics of microRNA-mRNA interactions. However, they cannot give us a complete collection of sites for a certain organism, since they fail to discover sites on mRNAs with little or no expression or when the corresponding microRNAs are missing in the given cell type. Computational target prediction based on principles learned from these collections has the potential to identify those left-out target sites.

In this work, several up-to-date, high-throughput collections of human microRNA target sites have been collected and processed in order to use them as training and test sets for a novel graph-based prediction model. Chapter 1 first introduces the reader to the complex biology of microRNAs, followed by a description of available methods for target identification, including the high-throughput-methods used to generate the collected data. The last two sections comprise the motivation and objectives for this thesis, followed by a description of the thesis structure.

1.1. microRNA biology

MicroRNAs (miRNAs) comprise a broad class of typically 20-23 nt single-stranded endogenous non-coding RNAs that regulate gene expression [1, 2]. They have been found in animals and plants, but not in fungi, which together with their distinct features in biogenesis, processing and mode of action suggests that miRNAs have evolved at least twice in the two eukaryotic lineages [3]. Some miRNAs can regulate the expression of up to hundreds of genes and one gene can hold up several miRNA binding sites, which can act in a cooperative manner in order to enhance the regulatory effect [1]. Although the impact of miRNA-mediated regulation on gene expression is usually modest, miRNAs are involved in the regulation of most genes [4, 5], thus directly affecting the majority of cellular pathways, from development to tumorigenesis [6].

Since this thesis is about miRNA target prediction in humans, the following descriptions focus on animal or mammal miRNA biology, thus omitting further explanations on plants. Subsection 1.1.1 illustrates the mammalian miRNA pathway from biogenesis to processing. The following two subsections 1.1.2 and 1.1.3 focus on miRNA mode of action, with the second one illustrating some recent findings that are likely to expand our current picture of miRNA functionality in the near future.

1.1.1. Biogenesis and processing

MiRNA biogenesis begins with the transcription of primary miRNA transcripts (pri-miRNAs), which are encoded either by miRNA genes or in the introns of mRNA genes [2]. Figure 1.1 illustrates the canonical biogenesis and processing pathway. It represents the standard maturation pathway for most miRNAs, although literature reports numerous exceptions along that way. For example, miRNA editing, differential trimming (isomir generation) or an alternative branch that escapes Dicer processing in the cytoplasm can take place for certain miRNAs [2, 3].

Transcription of pri-miRNAs results in single or, in the case of miRNA clusters, polycistronic miRNAs [3]. Following transcription, the so-called Microprocessor complex, consisting of endonuclease Drosha and RNA binding protein Pasha (DGCR8 or partner of Drosha), cleaves the pri-miRNA into an ~ 60 -nt-long stemloop structure termed pre-miRNA. After its export into the cytoplasm by the Exportin-5-Ran-GTP complex, further cleavage through interaction with endonuclease Dicer together with double-stranded RNA-binding protein TRBP finally results in a ~ 22 -nt-short, double-stranded RNA made up of two mature miRNAs. The generated duplex is then loaded onto an Argonaute (AGO) protein, where one mature miRNA strand is selected to act as the guide strand. This strand subsequently directs AGO as the center of the miRISC (microRNA-induced silencing complex) to its complementary mRNA target site, teaming up to exert targeted regulation. The second miRNA, also termed passenger strand, is released and degraded [2].

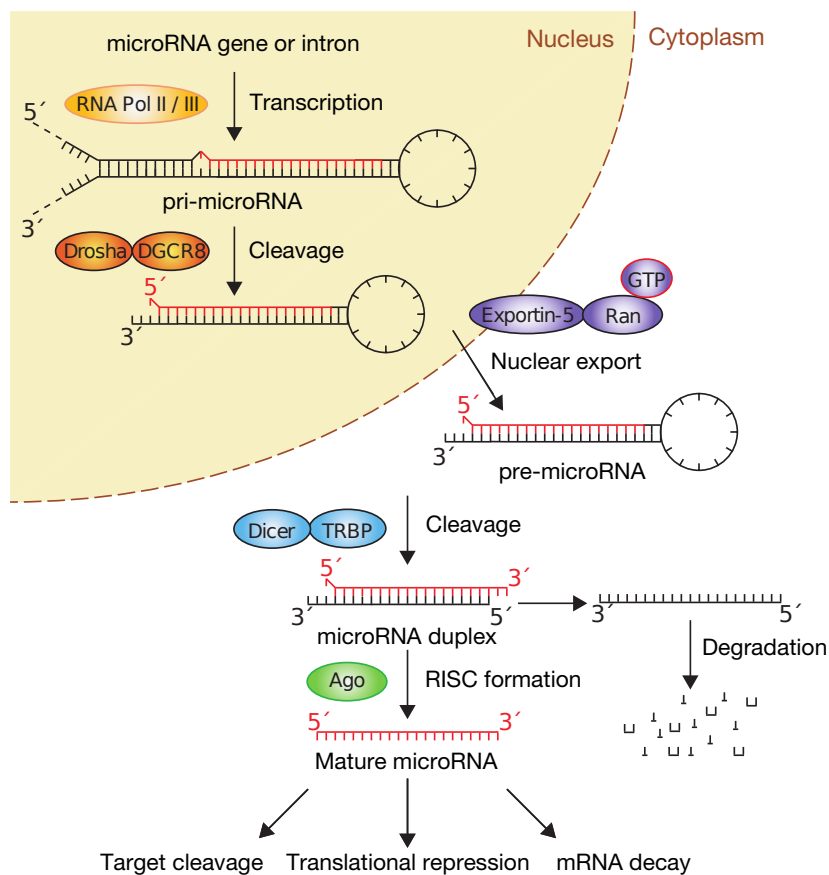


Figure 1.1.: Canonical view on animal miRNA biogenesis and processing. The miRNA gets transcribed from miRNA or mRNA (intronic) genes as pri-miRNA and cleaved by Drosha-DGCR8 Microprocessor complex to generate the pre-miRNA. It is then exported into the cytoplasm by Exportin-5-Ran-GTP complex, further cleaved by Dicer endonuclease, followed by its loading onto the miRISC complex, strand selection and miRNA guidance of AGO to its target mRNA. Binding and interaction with additional proteins finally results in the exhibition of its regulatory effect by either translational repression, mRNA decay or cleavage (figure modified after [2]).

1.1.2. Mode of action

In order to exert their regulatory role, miRNAs need to be bound by AGO proteins [7]. There are four distinct AGO proteins in humans that participate in miRNA-mediated regulation of gene expression, with AGO2 being the most abundant one [8]. In addition, AGO2 is unique among the others for its slicing activity, enabling cleavage of highly complementary miRNA target sites. The extent of cleavage in animals is unknown but expected to be rather uncommon, since there usually is only limited miRNA-target complementarity [1, 3]. Other than this difference, all four AGO proteins are reported to bind highly similar sets of transcripts, indicating substantial functional redundancy [9, 10]. Moreover, miRNA sorting to individual Argonautes seems to be mostly random, although there might be a sorting mechanism for at least some miRNAs [8, 11, 12].

After joining forces, miRNAs guide their associated AGO proteins to complementary mRNA target sites. One particularly important factor that stabilizes miRNA-target interactions involves strong base pairing between the target and the 5' end (positions 2 to 7)¹ of the miRNA, generally referred to as the *seed* region of the miRNA [1]. Such miRNA seeds also play a pivotal role in animal computational target prediction, since extensive base pairing between the seed and the target alone is often functionally sufficient (see Chapter 2 for details). Up to recently, miRNA target sites were thought to be located primarily in the 3' untranslated region (UTR) of the target mRNAs [1]. Lately however, there have been various reports about biologically functional target sites in the coding sequence (CDS), with smaller but still measurable effects on gene expression [10, 14, 15]. Apart from sequence complementarity and target-site location, there are many more target site features that influence the affinity of the site towards miRNA binding. Since these features are also considered in computational target prediction, a detailed description of them is given in Chapter 2.

Ultimately, miRNA binding results in changing the expression of the target, mediated by the interaction of AGO with various other proteins such as GW182 during translation [7]. Down-regulation of gene expression seems to be the predominant case, although reports do exist about stimulation of gene expression induced by miRNAs [16, 17, 18]. There is still an ongoing discussion about how down-regulation of expression is actually achieved, i.e. if translational repression precedes mRNA degradation or not [19]. Several recent publications, however, support the notion that translational repression is the initial event, followed by mRNA degradation [20, 21, 22].

1.1.3. Refining the biological model

Although the standard model of miRNA-directed mRNA regulation is well established, an increasing amount of scientific work outlines unexpected findings that will likely force some of the current notions about miRNAs to be modified in recent years. Besides the previously mentioned works about the up-regulation of gene expression through miRNAs [16, 17, 18], miRNA binding sites can also be found on a growing class of noncoding RNAs [23]. These RNAs are thought to act as miRNA sponges, which means that they compete with mRNA target sites for miRNA occupancy, thus adding an additional layer of gene regulation. Among these transcripts, some recently discovered circular RNAs (circRNAs) seem to act as particularly efficient sponges, most likely due to their natural resistance to exonucleolytic RNA decay [24].

Another surprising finding is the ability of miRNAs and AGO proteins to enter the nucleus [25, 26, 27, 28]. This way, miRNAs might additionally regulate nuclear RNAs while AGO and other RISC proteins have been shown to act as transcriptional coregulators by interacting with nuclear receptors. Moreover, there are reports that link AGO to other nuclear processes, such as chromatin modification, DNA repair and alternative

¹more recently also 2 to 8 or in general 1 to 8 [13].

splicing [7]. The same article also mentions several publications where AGO has been found to reach mRNA targets without miRNA guidance, nonetheless resulting in target repression. It is speculated that AGO recruitment to mRNAs can be achieved by RNA binding proteins solely, although the extent of this phenomenon is yet to be revealed.

1.2. Methods for microRNA target site identification

Studying miRNA-directed mRNA regulation requires the identification of miRNA-mRNA interaction partners, ideally together with precise target-site localization. Available methods can be categorized in computational, target-specific or high-throughput methods.

1.2.1. Computational methods

Computational methods apply the biological principles that direct miRNA-mRNA interactions in order to make reasonable predictions. These principles are derived from a growing amount of experimental data, resulting in increasingly sophisticated prediction programs. Computational prediction usually marks the first step in e.g. finding out whether a certain mRNA is regulated by miRNAs or not [29], followed by a target-specific method that experimentally verifies the predicted interaction (described in Section 1.2.2). Computational methods naturally excel in the rapid and low-cost identification of potential miRNA targets, as well as their ability to be integrated into large-scale studies e.g. combined with expression profiles, gene ontology or pathway annotations. On the other hand, currently available prediction programs are still limited regarding their predictive performance. For example, [30] compared the performances of 9 popular programs on high-throughput experimental data, resulting in less than 50 % sensitivity at best and about 50 % precision at 6 to 12 % sensitivity. The authors also showed that combining predictions from different programs is not advisable due to small overlap between predictions, which would result in the loss of many true-positive targets. Since various programs use various principles and because of their importance for understanding computational prediction, an in-depth review of them is given in Chapter 2.

1.2.2. Target-specific methods

Target-specific methods include experimental techniques for validating the interaction of one miRNA with an assumed target mRNA. The assumption can be based on a computationally predicted interaction, which needs to be experimentally validated. One common procedure uses target-reporter-gene constructs transiently transfected into cells,

combined with over-expression or down-regulation of the corresponding miRNA through miRNA or antisense oligonucleotide transfection [31]. The negative control can be the target sequence containing mutations in the target site, in order to exclude secondary effects and to localize the target site. Also, verification of miRNA target co-expression is usually performed for further validation. This method is favorable over other gene-specific methods in that it is possible to prove a direct interaction and validate the exact target site. However, there are also drawbacks: the use of non-physiological miRNA or target concentrations can lead to non-physiological interactions. Moreover, the target sequence is generally just the 3' UTR, thus risking the loss of important context. In addition, for loss-of-function studies using antisense oligonucleotides to silence endogenous miRNAs, the inhibitors might only be specific for a certain miRNA, but not for other miRNA family members with similar sequences. These shortcomings plus the inherent inability to study gene network effects can be addressed by high-throughput methods.

1.2.3. High-throughput methods

Owing to the advent of microarrays and particularly next-generation sequencing, numerous transcriptome-wide studies and techniques to identify miRNA-mRNA interactions have been published in recent years. Changes in transcriptome-wide mRNA levels upon miRNA transfection can be measured using microarrays, giving clues about regulatory networks [32]. Similarly, protein levels have been examined upon miRNA transfection and knockdown [4]. Although these methods yield valuable information about the impact of miRNA regulation and regulated gene networks, they cannot distinguish between direct and indirect miRNA targets. Furthermore, the same criticisms for over-expressing or down-regulating miRNAs apply as stated in Section 1.2.2.

To overcome these obstacles, a technique called HITS-CLIP (high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation) was successfully adapted to identify direct miRNA targets [33]. Briefly, AGO proteins get crosslinked to interacting mRNAs by UV radiation of cells, followed by AGO immunoprecipitation, reverse transcription of the crosslinked RNA fragments and cDNA deep sequencing. One particularly popular modification of the HITS-CLIP protocol termed PAR-CLIP (photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation) uses the photoreactive ribonucleoside analogue 4-thiouridine (4SU) in order to enhance crosslinking efficiency [10]. Their experimental procedures are illustrated in Figure 1.2. Importantly, both techniques do not require miRNA over-expression or down-regulation and can be performed either *in vitro* or *in vivo* (in case of HITS-CLIP). Moreover, PAR-CLIP frequently introduces T to C transitions at crosslinked sites, which can be exploited in subsequent mutational analysis to literally pinpoint the interaction region. So far, CLIP methods have been successfully applied in numerous studies, e.g. to identify the mRNA-bound proteome [34]. However, it has to be stated that one can not tell from CLIP data if the identified CLIP sites are functional, i.e. if they result in an effective change of mRNA or protein abundance. Moreover, reverse transcription efficiency lacks due to the short

polypeptide left after proteinase K treatment at the crosslinked site, resulting in substantially decreased recovery of binding sites [35]. This problem can be addressed by another modified CLIP approach called iCLIP [36].

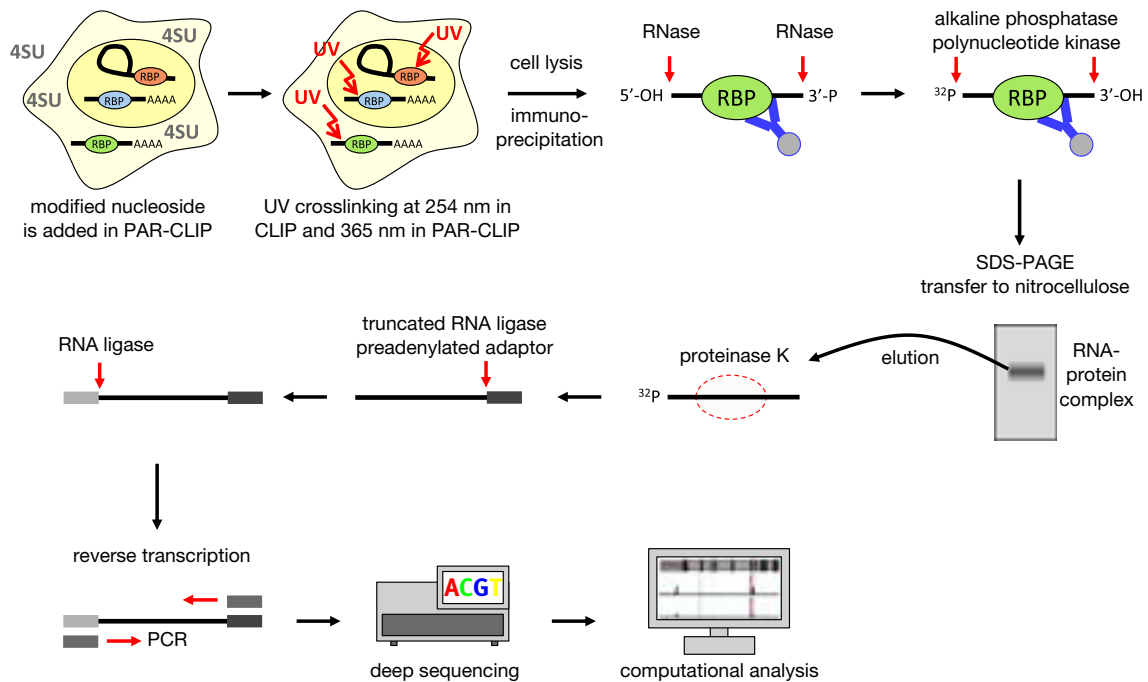


Figure 1.2.: Illustration of the HITS-CLIP procedure. Both the original HITS-CLIP (CLIP) and its modification PAR-CLIP apply the same principles in order to identify AGO (or other RBP) binding sites. Cells get UV radiated to obtain crosslinked AGO-mRNA complexes, followed by their immunoprecipitation, partial RNase digestion, radioactive labeling, Recovery by SDS-PAGE, transfer to nitrocellulose membrane to abolish loose RNA fragments, excision and proteinase K treatment to remove AGO and recovery of RNA segments. Segments then get reverse-transcribed and their cDNAs subjected to deep sequencing, followed by computational analysis. The main differences between the two techniques arise in the use of a modified nucleoside to enhance crosslinking (PAR-CLIP) as well as different UV wavelengths and RNases to trim the crosslinked RNA segments (figure modified after [37]).

In this work, four CLIP datasets from two publications have been utilized for training [10, 38]. In addition, a fifth dataset of miRNA target sites was used for testing and comparison, generated by a similar high-throughput technique called CLASH (crosslinking, ligation, and sequencing of hybrids) [39]. In contrast to CLIP, CLASH has the ability to identify a miRNA target site together with its corresponding miRNA, lending itself well e.g. as a test set for models trained on CLIP data. Identification of both target site and miRNA is accomplished by using an additional ligation step after UV crosslinking, AGO purification and partial RNase digestion. This results in joining of the miRNA-target duplexes crosslinked to AGO and the generation of so-called chimeric reads consisting of both miRNA and target site sequences. Verified miRNA-target pairs represent the desired data format for learning target site features and miRNA-mRNA

interaction principles, thus making CLASH a promising variation of the protocol. However, additional studies need to be conducted to compare performance and consistency of CLIP and CLASH regarding target site identification.

1.3. Motivation and objectives

Computational prediction of animal miRNA-target sites imposes a tough challenge on science, since complementarity of functional miRNA-target interactions is usually small, frequently just involving the seed region. This limited complementarity inevitably leads to a high number of false positive predictions, forcing programs to apply additional principles of miRNA targeting in order to increase specificity by filtering out false positives. However, it has been shown that popular prediction tools that rely on these principles perform far from optimal on experimental data, resulting in less than 50 % sensitivity at best and about 50 % precision with sensitivities from 6 to 12 % [30]. It is therefore of primary importance to refine targeting principles in order to increase performance. To achieve this goal, the analysis and utilization of recently published high-throughput datasets of miRNA target interactions for refinement was defined as the first of three major thesis objectives.

The second major objective deals with recent observations regarding the overlooked abundance of functional imperfect (noncanonical) seed interactions [13, 39, 40, 41, 42]. According to these observations, noncanonical seed interactions are widespread, with estimates from 15 % up to over 60 % of all seed interactions. However, current target prediction tools either ignore these interactions altogether or recognize only a small subset of them, while concentrating on perfect (canonical) or nearly perfect seed interactions, which results in decreased sensitivity. Objective number 2 was thus to combine a wide range of reported noncanonical and canonical seed interactions into a novel predictive model, which has not been done to this extent by any other available method.

The third major objective encompassed the construction of the predictive model. For this purpose, a novel graph-based machine learning model was adapted to the requirements of miRNA target prediction, and subsequently trained and tested with the compiled datasets. This way, the model's ability to discriminate between an increased number (due to including noncanonical interactions) of predicted positive and negative interactions was evaluated. An additional dataset with transcriptome-wide RNA binding protein (RBP) information was also added to the model and its influence on classification performance was tested. Summing up, the following thesis objectives were defined:

- Analyse the high-throughput data to understand their characteristics.
- Compile training and test datasets of miRNA-mRNA interactions.
- Include noncanonical seed interactions in positive and negative sets.

- Integrate various interaction site features into a novel graph model.
- Train the model by examining the different features and settings.
- Test the model on an independent test dataset.

The implementation of these objectives is detailed in Chapter 4. An overview of the thesis structure can be found in the next section.

1.4. Structure

Following the introduction to miRNA biology and available methods for miRNA target site identification in Chapter 1, Chapter 2 outlines known principles of miRNA target interactions which are applied in computational target prediction, together with mentioning some of the related popular prediction programs. Chapter 3 illustrates the graph-based machine learning approach utilized in this work, together with additional explanations necessary for understanding the topics of this thesis. Chapter 4 details individual steps in data acquisition, processing, mapping and generation of positive and negative sets, as well as generation of the extended model and implementation of the pipeline applied in model training and testing. Chapter 5 presents the results of data analysis, model extension, training and testing, and discusses them. Finally, Chapter 6 gives a summary on the obtained results and concludes the discussion with respect to future tasks and topics. In the appendix, further computational details on utilized tools are depicted which would have gone beyond the scope of the main chapters.

Principles of microRNA targeting

This chapter presents the miRNA target interaction principles and target site features which are utilized in computational miRNA target prediction. Over the last ten years, numerous methods have been developed that apply established, modified or new—mostly combined—principles and features in order to improve prediction of target sites. Some popular prediction programs will be mentioned while describing the principles and features which have been applied by them.

Despite their vast number, most algorithms commence by searching for mRNA segments complementary to the miRNA seed region (defined in 1.1.2) in order to generate an initial list of potential target sites [30]. The first section of this chapter will thus deal with sequence complementarity used as a predictive feature. Further sections describe thermodynamic stability, site accessibility, evolutionary conservation and additional approaches such as combinatorial, context- or site-specific principles.

2.1. Sequence complementarity

As opposed to plants, animal miRNA target interactions normally feature a limited amount of sequence complementarity between the miRNA and its target (described in Section 1.1.2). The miRNA seed region constitutes an exception in this case, since strong base pairing between the target and the seed is the most prominent feature of animal miRNA target interactions [1]. As expected, the seed region happens to be the evolutionary most-conserved sequence part of animal miRNAs. Moreover, miRNA seed

motifs (seed complements) are significantly enriched in regulated target sequences [32, 43]. Searching for seed motifs in mRNA sequences marks the first step taken by almost all popular miRNA prediction programs in order to find interaction sites, thus constituting the main prediction feature in animal miRNA target prediction [30].

The first programs which utilized the seed complementarity feature used a strict search, which means that contiguous base pairing of 7 or 8 nucleotides between the seed and the target is required for a match. More recently, however, it was discovered that these perfectly matching canonical seed sites are not the only functional seed interactions [44, 45, 46, 41, 42]. Imperfectly matching, so-called noncanonical seed interactions have been shown to be functional and widespread. These interactions usually contain 6 base pairs, including G:U wobble base pairs, single nucleotide bulges or mismatches. Seed search therefore became less strict, e.g. DIANA-MicroT-CDS [14], published in 2012, searches for initial hits which can contain a single G:U base pair or a single bulge or mismatch in the seed alignment after four consecutive Watson-Crick base pairs starting at seed positions 1 or 2. Regarding the seed search implementation in this thesis, an even less strict search was performed in order to incorporate more noncanonical interaction sites than currently available methods. All chosen noncanonical seed interactions have been taken from literature and are detailed in Chapter 4.

Despite the prevalent role of seed complementarity in target prediction, it has been shown that complementarity to the central and 3' region of miRNAs can also be important for functional targeting [47, 48]. Particularly, 3' end complementarity can supplement seed complementarity or compensate for the weaker pairing found in noncanonical seed interactions [1]. Also, 3' UTR motifs complementary to miRNA positions 12 to 17 (especially 13 to 16) seem to show a small amount of evolutionary conservation [48]. However, extensive sequence complementarity alone does not always conclude that the interaction is functional. Moreover, seed motifs, especially noncanonical, are extremely frequent in mRNA sequences. It is thus mandatory to filter out the false positive interactions generated by seed search / alignment, in order to increase specificity. Principles which can be applied for this job are described in the following sections.

2.2. Thermodynamic stability

Thermodynamic stability of the miRNA-mRNA duplex is commonly utilized by programs such as miRanda, Pictar or DIANA-microT-CDS [49, 50, 14] to extend the sequence complementarity measure. This stability is expressed in the minimum free energy of the duplex secondary structure and can e.g. be calculated by RNAhybrid [51]. That way, the interaction is evaluated by considering the stability of the duplex, using the calculated energy as a quality measure for the interaction. By setting a reasonable energy cutoff, hybrids with less-stable secondary structures are removed from the initial list of predicted interactions. Although thermodynamic stability is utilized in many prediction programs for ranking as well as removing false positive candidates, the fact

that animal miRNA target interactions rely on hybrids with partial complementarity also implicates that interactions do not need a particularly stable secondary structure in order to be functional. Indeed, it has been shown that ranking perfect 7mer seed containing sites by their hybridization energy does not perform significantly better than random ranking [52]. Besides, the impact of target secondary structure on hybrid formation is not considered. This impact, which can be approximated by measuring target site accessibility, is detailed in the next section.

2.3. Target site accessibility

Target site accessibility depicts an energy-based measure which essentially quantifies the potential of a given target site to be single stranded and thus accessible for miRISC binding. It is defined as the difference between the energy of the set of all structures and the energy of the set of structures where the target site is single stranded [53]. Since the computational costs of RNA secondary structure prediction increase at least quadratically with sequence length¹, the predicted target structure is usually limited to several hundred nucleotides. Correlation of 3'UTR target-site accessibility with repression strength and AGO2 cleavage efficiency has been shown in several studies [54, 55, 56], and the feature has also been incorporated into prediction programs, most notably PITA [54]. A second more simplistic way of measuring site accessibility can be obtained by checking the AU content in the up- and downstream vicinity of the target site. In agreement with the mentioned site accessibility studies, AU content was shown to be elevated around interaction sites [48] and is therefore also used as a prediction feature e.g. by TargetScan [5]. However, since contemporary RNA folding programs cannot yet consider RNA-protein interactions and their impact on secondary structure, ranking or filtering by site accessibility still has room for improvement [52].

2.4. Evolutionary conservation

Evolutionary conserved biological sequences indicate functional preservation upon selective pressure over the course of time. This conservation across species can be utilized for miRNA target prediction by searching for homologous, seed site containing mRNA segments. Indeed, many prediction algorithms use this feature to successfully reduce the number of false positive predictions, such as TargetScan, Pictar or EIMMo [43, 50, 57]. As stated in Section 2.1, seed sites have been shown to be evolutionary conserved and significantly enriched in regulated target sequences. It has also been reported that more than half of all protein-coding genes undergo evolutionary selection to maintain miRNA targeting [5], and that thousands of genes whose expression patterns overlap with miRNA

¹achieved by discarding multiloops and restriction of internal loop length [53].

expression patterns have evolved to selectively avoid target sites that match these miRNAs [58]. However, using strict conservation filtering also means losing many genuine target sites, since non-conserved mRNAs and miRNAs which evolved after the last considered segregation event cannot be detected. Furthermore, non-conserved sites have been reported to outnumber conserved sites 10 to 1 [58], while [40] found that 40 % of the miRNA target sites in the PAR-CLIP dataset [10] are non-conserved. It is thus advisable to not solely rely on the conservation feature for precision enhancement, although it constitutes a superior filter over site accessibility if conservation information is available [52, 59].

2.5. Additional principles

Apart from the features described so far, various additional principles have been applied in recent years in order to improve prediction performance. These principles e.g. comprise consideration of expression data, combined mode of action or the location of the target site, and will be presented in the following paragraphs.

Since miRNA target interactions require sufficient expression of both miRNA and target, the utilization of expression profiles in miRNA target prediction naturally lends itself to improve specificity. Due to the increased availability of miRNA and mRNA profiles, validated co-expression can be used as a feature to narrow down the search space of potential targets by preserving the biological context of miRNA-mRNA pairs. The feature is utilized e.g. by CoSMic [60], where both miRNA and mRNA expression data from the same samples are taken into account together with formerly described principles. This approach is clearly favorable if expression profiles exist, since many even high-scoring predictions without biological context can easily be filtered out. However, gene expression (miRNA and mRNA) is both tissue- and individual-specific [61, 62], which might result in the assumption of a biological context that does not exist in the examination object, if profiles are taken from public resources.

Concerning a combined mode of action, [48] discovered that canonical seed sites in close proximity¹ on the target mediate stronger repression than expected from the contribution of two single sites. The effect was observed both on sites close for different and identical miRNAs. Moreover, conservation analysis in human, mouse, rat and dog showed an enrichment of closely spaced coconserved sites. Several studies have confirmed this form of miRNA target site cooperativity [63] [64]², and it is also applied as a predictive feature e.g. by miRror [65].

Another important aspect in miRNA target prediction is the target site location on the mRNA. As mentioned in 1.1.2, it was originally assumed that miRNAs primarily

¹with an estimated intersite spacing of 8 to 40 nt.

²with refined spacing of 15 to 26 nt between 5' seed start positions.

interact with their target mRNAs by binding to complementary 3'UTR sites [1]. Most of the mentioned target prediction programs thus restrict their target search on 3'UTR sequences. Moreover, basically all known targeting principles have been discovered based on examinations of 3'UTR interactions. An increasing amount of reported target sites in the CDS has recently changed this view [46, 66, 67], as these sites are thought to have less pronounced but still measurable effects on gene expression. High-throughput datasets of miRNA target interactions [10, 33, 38, 39] report an abundance of target sites located in the CDS, which depending on the dataset even surpasses the number of 3'UTR sites¹. As a result, several prediction programs have been modified in order to include predictions in CDS sequences [14, 15], revealing additional binding characteristics. For example, seed complementarity seems to be more strict for CDS sites, resulting in stronger binding, and genes with shorter 3'UTRs tend to have significantly more CDS targets. Also, [68] showed that CDS sites are more effective in rapid translational repression than 3'UTR sites, and that different miRNA families seem to have preferences concerning CDS or 3'UTR targeting. Moreover, [48] reported that target sites near the ends of long 3'UTRs are more effective than sites in the central part or sites closer than 15 nt to the stop codon.

¹for example, [10] reports 50 % CDS, 46 % 3'UTR and 4 % 5'UTR.

The graph kernel model

This chapter contains the definitions and explanations necessary for understanding the graph-based machine learning model which was extended in the course of this thesis to perform miRNA target prediction. miRNA-target interactions feature both sequence and structural properties, which need to be properly encoded in order to utilize these interactions as training and test datasets in a machine learning environment. Since RNA sequences can fold into many different probable RNA structures, an efficient representation for capturing classes of highly probable structures will be described in the first section. Section 3.2 presents a suitable graph encoding for RNA secondary structures, followed by an explanation of the applied graph kernel technique, which extracts features from RNA secondary structure graphs. Section 3.3 deals with model training by learning from the extracted features and subsequent testing, using a support vector machine classifier. The last section describes commonly used classification performance measures.

3.1. Abstract shapes structure representation

RNA secondary structure prediction usually yields many different nearly optimal structures beside the minimum free energy structure, which is often not the exact biologically valid structure. In order to abstract from the plethora of probable structures to a set of assessable representatives, [69] introduced the abstract shapes approach. A shape is defined as an abstract representation of an RNA secondary structure [70]. The following example from the publication illustrates this for one sequence and two possible secondary structures in dot-bracket notation:

```
AUCGGCGCACAGGACAUCCUAGGUACAAGGCCGCCGCUU
..(((.(...((...)))..(((.....))))))..
..(((.....((...)))..(((.....)))..))..
```

The first out of five different abstraction levels or shape types begins with ignoring stacking and loop lengths, instead introducing underscores and square brackets for arbitrary lengths:

```
_[_[_[_]_]_]_
_[_[_]_]_]_
```

The fifth and most abstract shape type ignores unpaired regions altogether and merges nested helices. This way, two distinct RNA secondary structures can be represented by the same abstract shape:

```
[[] []]
```

The shape approach therefore allows abstraction from a huge set comprising all possible secondary structures of an RNA sequence to a much smaller set of structure classes represented by shapes. In addition, the structure with the minimum free energy inside a class (the shape representative or *shrep*) is chosen to represent each shape class. Shape analysis is implemented in the software package RNASHapes, which e.g. allows output of the most probable shapes represented by their shreps or all shapes inside a given energy range.

3.2. gSpan graph encoding

In order to utilize the graph kernel for extracting combined sequence and structure features from the shreps, the shape representatives have to be converted into graphs first. Graphs are well suited for encoding RNA secondary structures, since graph vertices and edges can perfectly resemble the nucleotides and backbone or base-pair bonds of an RNA structure. Moreover, efficient methods for graph processing are available. In this thesis, a graph kernel method is utilized which accepts graphs in the gSpan (graph-based Substructure pattern mining) format [71] as input. In gSpan format, all the graph information is stored in a plain text file, where each graph starts with a header row that begins with "*t #*", followed by a whitespace and the graph identifier string (see figure 3.1 A). After the header, the vertices get listed line by line with the format "*v vertexID vertexLabel*", followed by a listing of the edges with the format "*e vertexID1 vertexID2 edgeLabel*". In case of RNA secondary structure graph encoding, these edges can be backbone edges or base-pair edges.

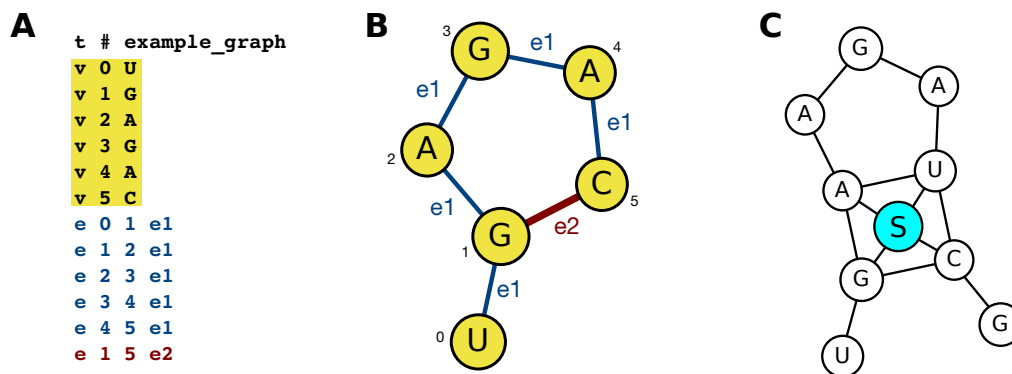


Figure 3.1.: gSpan graph encoding. **A:** A short RNA structure encoded in gSpan format. The second vertex column contains the numerical vertex IDs (counting up from 0) followed by the vertex label. Edge rows include the two vertex IDs they connect, plus an edge label (here "e1" for backbone edges and "e2" for base-pair edges). **B:** The same RNA structure visualized as a graph. **C:** Additional encoding of stacking information. The stacking vertex shares an edge with each of the four vertices that form the two stacking base pairs.

One addition to the RNA structure graph annotation can be the inclusion of stacking information (figure 3.1 C), as done in GraphClust [72]. Here, one vertex per stacking event is inserted, as well as the four edges that connect the vertex with the four vertices that form the two stacking base pairs. A corresponding extension for miRNA-target stacking events was added in this work, along with other extensions to capture the inherent information content of miRNA-target interactions (detailed in chapter 4).

Among these other extensions, three have also been applied in [73]. The first one transforms the undirected graph used in GraphClust into a directed graph. The direction of the graph is given by the edge direction (order of the vertex IDs in the edge rows) and is defined such that the 5' to 3' direction used for nucleic acid sequences holds for the graph as well. The second extension adds *abstract structure information* to the graph by using a hypergraph approach (figure 3.2A). For every secondary structure element, a hyperedge vertex and a vertex that represents the element is created. The hyperedge vertex then connects all the nucleotide vertices belonging to its element to the respective element vertex. The element vertices themselves are connected to each other the same way their structural elements are arranged in the RNA secondary structure (figure 3.2A, right side). This way, additional information about neighbouring substructures and belongings of features to certain substructures can be added to the graph. The third and last extension introduces the notion of *viewpoints* (Figure 3.2 B). This concept allows restricting feature extraction to a certain area of the graph. Since it is based on two parameters that control feature extraction done by the graph kernel, it will be explained in the next section together with the parameters and the principle functioning of the graph kernel.

Figure 3.2.: Abstract structure and viewpoint graph extension. **A:** Adding abstract structure information to the graph. Nucleotide vertices get linked to their respective secondary structure elements (stem, bulge or loops) using a hypergraph approach (figure modified after [73]). **B:** Adding viewpoint vertices to the graph. Using the viewpoint option, feature decomposition gets rooted to the viewpoint area, and extends only as far as the marked feature area.

3.3. Graph kernel feature decomposition

In order to learn predictive models from graphs, efficient graph processing methods are required. Graph kernels can be described as functions which measure the similarity between two graphs. Since similarity information is sufficient for binary classification of instances, graph kernels can be utilized for this task. In this thesis, the Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) [74] was used, since it efficiently computes the similarity between two graphs in linear time. The main concept behind NSPDK is that it extracts graph features by decomposing the graph into pairs of neighbouring subgraphs, controlled by two parameters (distance d and radius r). Radius r marks the maximum size of the subgraph, while distance d denotes the maximum distance between the roots of the two subgraphs, determined by the shortest path between them. Figure 3.3 gives an example for two different sets of parameter values. For every unique pair-of-subgraphs feature in the graph, the graph kernel subsequently stores its number of occurrences in a feature vector. Note that if $r > 1$, all possible subgraphs up to size r will also be extracted (the same applies for all possible pairs of subgraphs up to distance d if $d > 1$).

In order to gain additional control over feature extraction, the viewpoint graph extension mentioned in Section 3.2 was included to allow the restriction of one of the subgraph roots to a specified area of the graph (dark-green area in Figure 3.2 B). The dark-green vertices are denoted as viewpoint vertices (or simply viewpoints), whereas the feature area (all green-colored vertices) marks the region of feature extraction. The range of

the light-green area is defined by $d + r$, since one root can still be outside the viewpoint area. In the example graph, one valid value combination for r and d would therefore be $r = 1$ and $d = 3$.

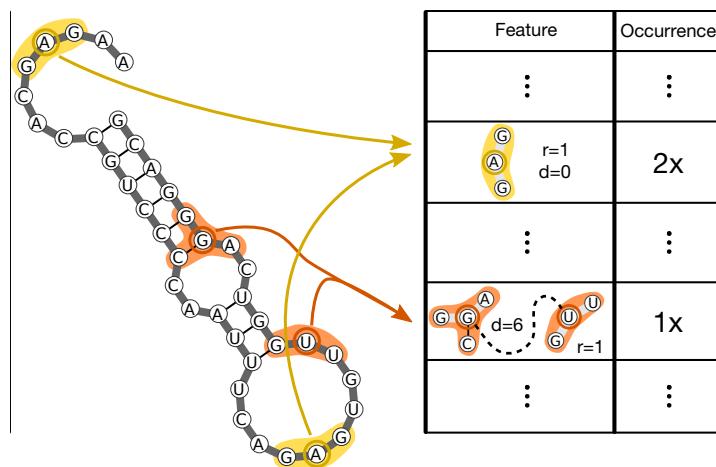


Figure 3.3.: Graph kernel feature decomposition. The NSPD kernel decomposes the graph into pairs of neighborhood subgraphs (based on the choice of values for parameters d and r) and stores the number of their occurrences in a feature vector (figure modified after [73]).

During the decomposition phase, one feature vector for every input graph is created and subsequently stored together with its corresponding classification or class label. In the case of binary classification, two class labels exist (e.g. "1" for positive instances and "-1" for negative instances). The generated feature vectors and the class labels are then analyzed by a suitable classifier in order to train the predictive model.

3.4. Model training and evaluation

Model training in this thesis was accomplished by using a Support Vector Machine (SVM) classifier. In general, miRNA target site prediction can be described as a supervised learning / classification task, where the classifier uses the presented positive and negative instances (true and false interactions) to train a predictive model. In the case of SVM classifiers, the instances of the two sets are represented as vectors in a multi-dimensional vector space. The classifier then tries to place a hyperplane into the space such that the margin between the nearest vectors (support vectors) of the two sets gets maximized. The resulting model then categorizes new, unseen instances according to the determined hyperplane.

In order to measure the ability of the model to correctly categorize new instances, error rates are commonly estimated by using a *cross-validation* technique. For example, in k -fold cross-validation, the training data is divided into k parts of the same size. The model is then trained on $k - 1$ parts and tested on the remaining one part. This procedure

is repeated k times, and the average performance measures are taken for evaluating model performance. It is important to note that when dividing the data into test and training datasets, any overlap between the two sets has to be avoided. This is because a classifier normally performs better on instances it was trained on, resulting in falsely high performance estimates.

3.5. Performance measures

Working with a binary classifier, the classification results are usually summarized in a confusion matrix (Figure 3.4). Based on the four entries in the matrix, all common classification performance measures such as sensitivity and specificity can be derived. These measures will be described in the following together with their equations¹.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Figure 3.4.: Confusion matrix for binary classification. The two class labels are "negative" and "positive". Entries a and b denote the number of correct and incorrect predictions, given that the instance is negative. Entries c and d denote the number of correct and incorrect predictions in case of a positive instance.

The true positive rate (TPR) (also termed *sensitivity*) describes the proportion of positive instances correctly classified by the model, and is calculated as follows:

$$TPR = \frac{d}{c + d} \tag{3.1}$$

The false positive rate (FPR) denotes the proportion of negative instances incorrectly classified as positive, given by the equation:

$$FPR = \frac{b}{a + b} \tag{3.2}$$

The true negative rate (TNR) (also termed *specificity*) describes the proportion of negative instances correctly classified:

$$TNR = \frac{a}{a + b} \tag{3.3}$$

¹taken from <http://www2.cs.uregina.ca/dbd/cs831/index.html>

The false negative rate (FNR) describes the proportion of negative instances incorrectly classified:

$$FNR = \frac{c}{c + d} \quad (3.4)$$

The *accuracy* (AC) denotes the proportion of the total number of predictions that are correctly predicted by the model:

$$AC = \frac{a + d}{a + b + c + d} \quad (3.5)$$

Finally, the *precision* (P) describes the proportion of predicted positive instances correctly classified, given by the formula:

$$P = \frac{d}{b + d} \quad (3.6)$$

A popular way of visualizing classification performance is the utilization of a *receiver operating characteristic* (ROC) curve. The x-axis of a ROC curve denotes the true positive rate (sensitivity), while the y-axis denotes the false positive rate (1 - specificity). In order to create a curve, model performance has to be measured repeatedly while adjusting the threshold for classifying an instance positive or negative. A commonly used performance measure concerning ROC is the *area under the ROC curve* (AUROC or AUC). It is calculated by the curve integral and was also used for evaluating the models in this thesis. The AUROC can be defined as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [75]. An AUROC of 0.5 (straight line with a slope of 1) means that the model performs equally to random guessing, while an AUROC of 1 denotes the perfect classifier (100 % sensitivity, no false positive predictions). Although there exists no definition about the quality of a certain AUROC, a value of ≥ 0.9 is often considered to be excellent, a value of 0.8 – 0.9 good, and a value of 0.7 – 0.8 still fair¹.

¹<http://gim.unmc.edu/dxtests/roc3.htm>

This chapter details the individual steps taken in order to obtain the results described in chapter 5. The first part deals with data pre-processing and initial analysis. Five high-throughput miRNA-target interaction datasets have been obtained and processed, and will be denoted as AGO1-4 PAR-CLIP, AGO2 CLIP, AGO2 PAR-CLIP, AGO2 PAR-CLIP-MNase and AGO1 CLASH in following sections. Section 4.2 describes the search for seed motifs as the first step of computational prediction, followed by IntaRNA hybrid prediction. Section 4.3 details gSpan graph generation and addition of miRNA-target interaction information. Section 4.4 depicts the construction of negative and positive datasets, pointing out similarities and exceptions in the creation of both sets. Finally, section 4.5 describes the computational pipeline that applies the described steps and logically connects them. Throughout this chapter, references to Appendix A will be given, which denotes some of the described concepts in greater detail.

4.1. Data pre-processing and analysis

The following sections detail the acquisition, pre-processing and initial analysis of the data used in this thesis. Notably, all five high-throughput interaction experiments as well as the RBP mRNA occupancy study used HEK293 (human embryonic kidney) cells to generate their results.

4.1.1. mRNA and microRNA sequences

The most recent collection (April 2013, 34038 mRNAs) of human RefSeq mRNA sequences (human reference genome version 19) was downloaded from UCSC¹. The NCBI RefSeq (Reference Sequence) transcript dataset provides a repository of non-redundant, well-annotated and daily-updated sequences. Sequence features such as exon and CDS annotations were taken from GenBank, using the BioPerl NCBI GenBank database interface (see Appendix A.1.1). Regarding the miRNA sequences, the most recent collection (April 2013, release 19) of human mature miRNA sequences was obtained from miRBase².

4.1.2. AGO1-4 PAR-CLIP dataset

The first PAR-CLIP dataset (AGO1-4 PAR-CLIP) is provided in the supplementary data (table S4) of the publication [10]. The spreadsheet file comprises 17319 combined AGO1-4 miRNA-target interaction sites, along with quality scores, genomic coordinates and sequences of the target sites. The sites were obtained from sequence reads (with a minimum of 5 reads and 20 % T to C transitions) aligned to the human genome (hg18) and extended to a length of 41 nt per site. Each cluster is centered on the predominant T-C transition (described in 1.2.3) in its sequence reads (position 21), marking the RNA-protein crosslink position and thus pinpointing the location of the miRNA-target interaction.

To determine the mRNA positions of the 17319 interactions, the 41-nt sequences were aligned to a locally set-up BLAST database of RefSeq hg19 transcripts (Appendix A.1.2). BLAST alignment yielded 14317 full mRNA hits (plus 1211 sites with partial, 1583 with no, and 208 with non-coding RNA hits). Two interactions were removed due to deletions in the target sequence compared to the reference mRNA sequence. In case of multiple full hits, the hit with the longest mRNA was chosen, which resulted in 5843 site-containing RefSeq mRNAs. 80 % of these mRNAs contain less than 4 target sites, with the highest occupancy being 33 target sites for one mRNA. Interestingly, several members of the miRNA targeting machinery, e.g. AGO1-4 or TNRC6A-C, appear among the transcripts with multiple binding sites (see Section 5.1.3). In order to map the sites to mRNA regions or exons, sites were converted into BED format and intersects were computed with BEDTools (Appendix A.1.3). Positions 21 of each site mapped to mRNA regions resulted in 2,2 % 5'UTR, 50,8 % CDS and 47,0 % 3'UTR occupancy.

For miRNA expression data, the list in supplementary table S5 (HEK_293_lysate) was utilized. Table S7 has the list of the top 100 expressed miRNAs, which was applied in various publications (e.g. [14, 59]). However, it contains some erroneous miRNA sequences (41 with wrong start or end nucleotides) and miRNAs that do not appear in

¹<ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/refMrna.fa.gz>

²<ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz>

the expression profile, and was thus rejected in favour of the table S5 expression profile list (as recommended by the author). Since an old miRBase release was used, the mature miRNA IDs had to be updated via sequence comparisons to the new database. In order to reduce redundancies, mature miRNA sequences with identical 2-8 seed sequence were summarized to families, as e.g. done in [38], and only the highest expressed family miRNAs (one miRNA represents one family) were taken into account.

4.1.3. AGO2 CLIP and PAR-CLIP datasets

A second publication [38] supplied three additional CLIP datasets (AGO2 CLIP, AGO2 PAR-CLIP, AGO2 PAR-CLIP_MNase), which were obtained from Gene Expression Omnibus (GEO) (GEO accession-ID: GSE28865). Out of the two replicates for each experiment, replicate A datasets were chosen since these showed more consistent results in the publication (GEO-IDs: GSM714642, GSM714644, GSM714646). The three dataset FASTA files contain 54905, 91362 and 44497 40-nt-long target site sequences, together with read counts and mapped transcript-IDs. In contrast to the AGO1-4 PAR-CLIP dataset, read coverage was used to center the sites. Also, no pre-filtering was applied by the authors, which explains the large numbers of interaction sites. For each CLIP site, Read coverage both in the foreground (CLIP sample) and the background (RNA-seq) is given. Thus, site enrichment can be calculated (foreground divided by background), which, according to the authors, should be favored over read coverage. As with the AGO1-4 PAR-CLIP data, the sites were first mapped to hg19 refSeq transcripts using local BLAST (Appendix A.1.2). This led to 54388, 90427 and 44113 full hits, together covering 9834 refSeq mRNA transcripts. Mapping of site positions 20 to mRNA regions yielded 1,5 % 5'UTR, 29,6 % CDS and 68,9 % 3'UTR occupancy for AGO2 CLIP (AGO2 PAR-CLIP: 1,6 %, 33,8 %, 64,6 %, AGO2 PAR-CLIP-MNase: 2,9 %, 38,8 %, 58,3 %).

As suggested, the miRNA expression profile was taken from the AGO2 PAR-CLIP-MNase sample and downloaded from the author-hosted CLIPZ server¹. Of the top 100 miRNAs, 85 also appear in the AGO1-4 PAR-CLIP top 100 miRNA expression list. Analogous to AGO1-4 PAR-CLIP, only the best family miRNA was taken.

4.1.4. AGO1 CLASH dataset

Concerning the CLASH dataset (AGO1 CLASH) [39], supplementary table S1 was taken, which contains 18514 miRNA-target interactions. An interaction is defined as a chimeric read, with both miRNA and mRNA fragments ligated. Site quality scores are given as the number of chimeric reads per site, as well as the number of experiments in which the chimeric read was found. The file also contains the information about which miRNA belongs to each target site, which was used for expression ranking due to the lack of

¹<http://www.clipz.unibas.ch/>

additional expression data. Site lengths are not normalized and vary from 18 up to 119 nt, with most sites having a length of 43–49 nt. BLAST Mapping was thus performed with a less strict e-value cutoff ($E = 0.001$ instead of $E = 0.00001$ for CLIP mapping), in order to capture the shorter sites. Transcript mapping resulted in 17938 full hits on 6969 mRNA transcripts, and subsequent region mapping yielded 4,4 % 5'UTR, 61,5 % CDS and 34,1 % 3'UTR region occupancy. Of the top 100 CLASH miRNAs, 54 and 56 can be found in the AGO1-4 PAR-CLIP and AGO2 top 100 miRNA list, respectively.

4.1.5. RNA-binding protein mRNA occupancy dataset

Essentially, the RBP mRNA occupancy dataset [34] includes millions of RBP binding site positions across the transcriptome. It was obtained from GEO (GEO-ID: GSE38355), where four supplementary files can be found in the repository, of which the two "consensus_TC" files were taken. These files contain 4740558 genomic protein crosslink coordinates (hg18) in BEDGRAPH format for the minus and plus strand, originating from two profiling libraries. In each library, the authors demanded that the number of T-C transitions at the genomic crosslink position should be at least 2, otherwise the position was removed. T-C transitions are common to the PAR-CLIP protocol (see Section 1.2.3), which was used in this study. Importantly, the dataset only contains crosslink coordinates, while the identity of the crosslinked RBP is unknown. In order to map the coordinates on mRNAs, a table containing exon coordinates of all human RefSeq genes (hg19) was downloaded from UCSC, using its table browser¹. For conversion of genomic coordinates from hg18 to hg19, the liftOver tool was used (Appendix A.1.4). Since some genomic coordinates erroneously appeared twice in the tables, entries with lower average T-C count were removed.

After mapping the T-C positions to genomic exon regions, the transcriptomic coordinates of the T-C positions had to be calculated. Some RefSeq transcript IDs with mapped target sites appeared twice in the downloaded RefSeq genes table. Here, the longer transcript version was taken. Also, some transcript IDs were not present in the table. In case of minus strand transcripts, it is important to note that the exon order is reversed in the RefSeq genes table, with the last mRNA exon start coordinates appearing first. Moreover, in the case of minus strand mRNAs, the reverse complement has to be taken in order to correctly calculate the T-C positions. Additionally, when dealing with BED files, one needs to remember that the first BED coordinate is zero-based (i.e. sequence position one is denoted as zero), while the second coordinate is one-based. RefSeq transcripts frequently contain short poly-A tails, but their exon coordinates do not. This has to be considered as well in the mapping phase. Finally, T-C position mapping to target-site-containing mRNAs resulted in 2738767 T-C positions, distributed across 10672 mRNAs.

¹<http://genome.ucsc.edu/cgi-bin/hgTables>

4.2. Seed scanning and hybrid prediction

As described in Chapter 2, virtually all miRNA target site prediction programs initially search for seed motifs in mRNA sequences, prior to applying more elaborate techniques in order to improve predictive performance. In this thesis, seed motif search was accomplished by utilizing sets of regular expressions in Perl to detect various reported noncanonical and canonical seed sites. Subsequent hybridization of the miRNA to the seed-hit containing segment was achieved by IntaRNA (Appendix A.2.2). Section 4.2.1 presents the incorporated seed types along with experimental evidence, Section 4.2.2 describes the seed scanning procedure and Section 4.2.3 details IntaRNA hybrid prediction.

4.2.1. Incorporated seed types

Although reports of experimentally verified noncanonical seed interactions date back well into the 1990s [76], only recently the substantial extent of noncanonical targeting has been uncovered, mostly due to results and analyses of high-throughput experiments [39, 40, 41, 42]. These publications report various noncanonical seed sites with 6 base pairs, as well as G:U base pairs, an additional bulge or a mismatch. Beside high-throughput studies, experimentally verified noncanonical target interactions have been reported by various studies. Table 4.1 lists some studies, without intending to be exhaustive.

Table 4.1.: Experimentally verified properties of noncanonical seed interactions. The listed studies utilized target-specific methods such as reporter gene assays (described in 1.2.2) in order to approve the functionality of the interaction.

Seed property	References
G:U base pairs	[42, 45, 46, 77]
Bulges in miRNA	[76, 77]
Bulges in mRNA	[41, 46]
Mismatches	[42, 46]

Apart from the mentioned studies, computational simulation of the AGO-miRNA-mRNA ternary complex was conducted in order to measure the structural stability of the complex while testing different seed interactions [78]. It was concluded that seed interactions containing multiple G:U base pairs as well as single bulges at several seed positions in both sequences do not affect overall complex stability.

Drawing from these findings, the following seed properties were incorporated into seed motif search such that seed sites which feature these properties will be recognized by the seed scanner:

- The seed has to contain at least 6 base pairs.

- Arbitrary number of G:U base pairs in the seed.
- One single nt bulge between positions 2 and 8 on the target.
- One single nt bulge between positions 3 and 8 on the miRNA.
- One mismatch at any position of the seed.

The last three properties are mutually exclusive, which means that if e.g. the seed site contains a mismatch, it must not contain an additional bulge in both seed sequences. In this thesis, the seed region was defined as the first eight 5' nucleotides of the miRNA sequence.

4.2.2. Regular expression seed scanning

Regular expressions which recognize seed motifs with described properties (Section 4.2.1) were constructed in Perl (see Appendix A.2.1). For each miRNA seed, its respective regular expressions were extracted and subsequently used together in a pattern matching operation to scan mRNA sequences for the seed motif. In order to prioritize seed hits that represent stronger seed sites, a four-step seed search was conducted. First, the scanner looked for contiguous 8mer seed hits (including G:U base pairs), followed by contiguous 7mer hits, contiguous 6mer hits, and finally all leftover 6mer seed hits. The results were filtered using `intersectBed` (Appendix A.1.3) such that overlapping, less strong seed hits were discarded. Since regular expression scanning is very fast, this approach does not impose any speed problems and helps to reduce the high numbers of initial seed hits due to the defined loose seed constraints.

4.2.3. IntaRNA hybrid prediction

IntaRNA (Appendix A.2.2) was utilized to predict the hybrid structure as well as the minimum free energy of the miRNA-target interaction. In order to capture the described seed properties (Section 4.2.1) during hybrid prediction, IntaRNA was set to search for a seed with a minimum of 6 base pairs and 2 as the maximum number of unpaired seed bases in both sequences. IntaRNA allows the definition of a seed region both for the miRNA and for the target (in case of the beta version described in A.2.2). The target seed hit region obtained from seed scanning was used as the target seed region for IntaRNA. Also, accessibility was disabled for the energy calculation in order to increase speed. In case of several IntaRNA hits, only the minimum free energy hybrid was considered. For easy evaluation of the calculated hybrid, an additional Perl script (`intarna_miRNA_target_analysis.pl`) was utilized which adds Watson-Crick / G:U base pair, bulge and mismatch annotations to the IntaRNA output hybrid.

4.3. Graph generation and extensions

Proceeding hybrid prediction by IntaRNA, the hybrid features were evaluated and converted into gSpan graph annotations (introduced in 3.2). Section 4.3.1 describes the generation of the gSpan file based on a sequence segment that includes the miRNA target interaction site. Section 4.3.2 details the miRNA graph extensions inserted into the gSpan file in order to capture the sequence and structural features of miRNA target interactions.

4.3.1. gSpan graph generation

Graph generation involves extraction of the sequence segment containing the predicted miRNA-target hybrid, shrep calculation of the segment (described in 3.1), and conversion of the shreps into the gSpan format. First, the segment was extracted based on the interaction region of the miRNA with its target. The interaction region ranges from the first to the last base pair of the hybrid, plus the miRNA overlapping nucleotides at both ends. The middle position of the interaction is then taken, and extended 150 nt on both sides. The resulting segments thus typically had a length of 301 nt, with the exception of target sites near the mRNA borders. In this case, the sequence up to the end was extracted. The segment was then converted into FASTA format and committed to an existing Perl script (`fasta2shrep_gspan.pl`) which executes shrep calculation and gSpan generation. Set parameters for the script are detailed in Appendix A.2.3. In short, shreps from the top three shapes were included, as well as abstract shapes, the unstructured segment and viewpoints that span the interaction region. The script outputs a graph file (`.gspan`), containing four subgraph sections: One section for the unstructured segment and three sections each including (in addition to the sequence) one shrep structure together with its abstract structure information. The viewpoint annotation was included in the sequence and structure sections.

4.3.2. miRNA graph extensions

Based on the generated gSpan file described in section 4.3.1, miRNA structure and sequence features were added to the graph by inserting vertices and edges into the four subgraph sections or creating new graph sections. Figure 4.1 A-E conceptually visualize the introduced graph extensions.

Starting with the interaction region (4.1 A), hybrid information was added to all four sections (4.1 B-C). Additionally, several ways to encode the interaction information were tested, and included as separate i-sections into the graph. Every such section contained vertices for the target segment and the miRNA, as well as backbone edges and base pair edges. Each i-section differs from the other i-sections in the designated labels for the

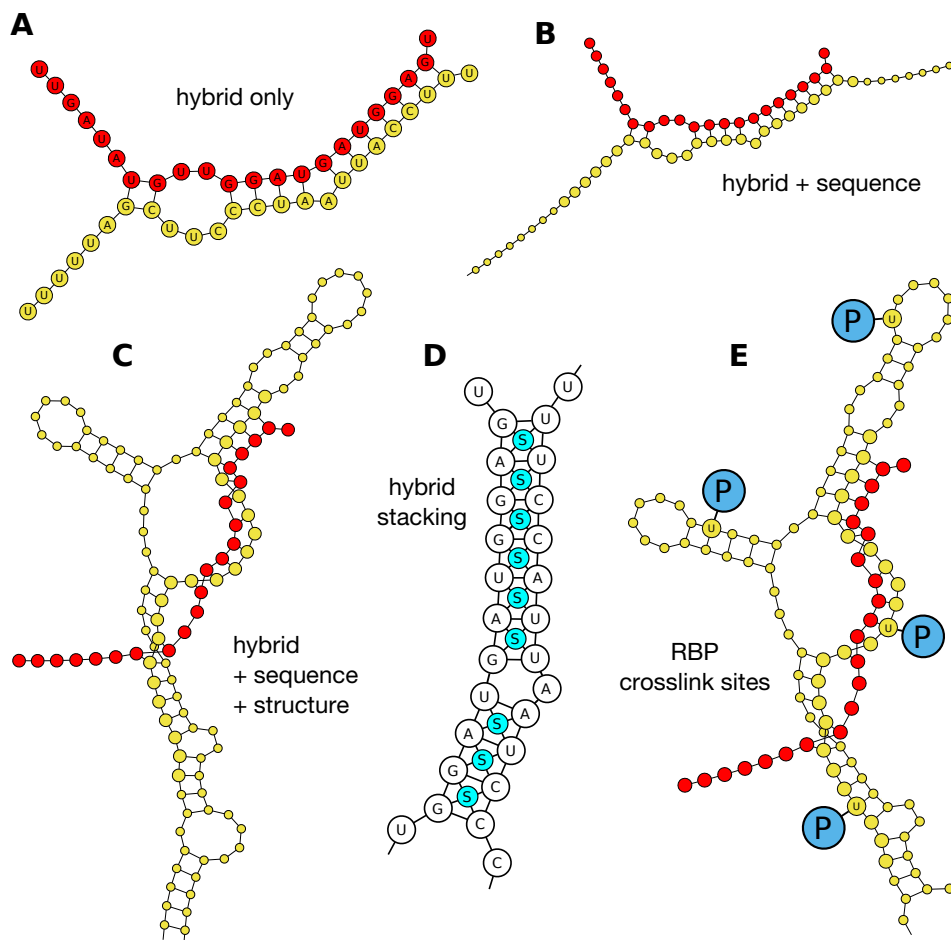


Figure 4.1.: Graph extensions for miRNA-target interactions. **A:** Hybrid interaction with miRNA (red) and target segment (yellow). **B:** The extracted target sequence with hybrid information. **C:** Target secondary structure with sequence and hybrid information. **D:** Adding stacking information to hybrid stacks. **E:** Adding RNA binding protein crosslink information to the target structure.

miRNA and mRNA vertices. Table 4.2 lists the vertex label contents of the different interaction sections. The predictive performance of each of the interaction sections was tested (see Results 5.2.1), and the best performing sections' labeling was used as labeling for the miRNA vertices in the four main sections.

The next step involved incorporation of stacking information into the miRNA-target hybrids (4.1 D), which can be found in the sequence section, the three structure sections and the remaining interaction section that showed the best performance. For every stacking event, one vertex and four edges was inserted, such that the four edges connect the stacking vertex with the miRNA and mRNA vertices that form the stack.

In the last step, RNA binding protein information was added (4.1 E). As described in Section 4.1.5, the information was made available in form of transcript coordinates. These transcript coordinates had to be converted into segment coordinates, and were

Table 4.2.: Vertex labels of the five included interaction sections. The "-" denotes that the same label was used for every vertex. In i4, miRNA nucleotides 1–8 get labeled seed, while the other miRNA nucleotides get labeled non-seed.

Interaction section	miRNA vertex label	mRNA vertex label
i1	Nucleotide and position	Nucleotide
i2	Nucleotide	Nucleotide
i3	Position	-
i4	Seed or non-seed	-
i5	-	-

inserted as protein vertices into an additional protein section after the interaction section for each graph. For each connection between an mRNA segment vertex (present in all four sequence and structure sections) and the protein vertex, an edge was inserted into the protein section. Concerning the protein vertex labels, three categories were chosen, depending on the average number of T-C transitions of the protein crosslink. Numbers 2–5.5 were labeled "low" (constituting 60 % of all crosslinks), numbers 6–21.5 were labeled "mid" (representing 35 % of all crosslinks), and numbers 22–1239 were labeled "high" (constituting 5 % of all crosslinks).

4.4. Construction of test and training datasets

The following sections describe the construction of positive and negative datasets for model training and testing. When constructing sets for model training, it is important to not introduce any biases that could be taken into account by the model in order to distinguish between positive and negative instances. Section 4.4.1 details the process of positive set generation, from seed scanning to the actual model training. Section 4.4.2 focuses on the specialities of negative set generation.

4.4.1. Positive interactions

Construction of the positive sets began with regular expression seed scanning (Section 4.2.2) the high-throughput miRNA-target sequences, analogous to negative set construction which started with seed scanning of the mRNA sequences. In case of CLIP sites, the 40 to 41 nt target sequences were searched for seed motifs¹ of the top-expressed miRNAs, while for CLASH the miRNA for each target sequence had already been identified. Concerning the AGO1-4 PAR-CLIP dataset, the seed search however did not encompass the whole 41 nt region. Rather, search was restricted to positions 20–30,

¹continuing at position +1 after a hit was found.

which showed to be enriched in seed motifs of highly expressed miRNAs [10]. Once a hit was found, IntaRNA (Section 4.2.3) predicted the hybrid, which then was analysed. Analysis involved extraction of interaction statistics from the hybrid and storing the statistics into tables. These tables were subsequently used for filtering (together with the negative sets) and as input tables for gSpan generation.

In the filtering phase, optional filtering could be applied in the form of stricter energy filtering or filtering based on the quality of the target sites. This quality was determined by either using site read counts (AGO1-4 PAR-CLIP, Section 4.1.2) or site enrichment (AGO2 datasets, Section 4.1.4). In order to use both site scores together in a merged CLIP set, scores were normalized by giving the highest score of the set a value of 1 and then transform the other scores proportionally to the highest score. Due to the small number of unique chimeric read counts, no site quality filtering was applied for AGO1 CLASH. Opposed to optional filtering, predicted interactions that did not meet the following criteria were always filtered out:

- A hybrid minimum free energy of < -4 kcal/mol.
- A maximum hybrid bulge size of ≤ 12 nt.
- Presence of protein crosslinks on the target interaction mRNA.

Also, in case of six base pairs and several G:U base pairs, a specific amount of compensatory 3' pairing (see Section 2.1) was required, depending on the number of G:U pairs. For example, in the case of 6 base pairs and ≥ 3 G:U pairs in the seed, positions 12–17 had to have at least 4 base pairs or positions 18–19 had to be paired. This was demanded since [13] showed that positions 18–19 are frequently paired in the AGO1-4 PAR-CLIP dataset. For the remaining cases with 6 base pairs and $<$ G:U base pairs, the number of required base pairs at positions 12–17 was stepwise reduced.

Proceeding filtering and gSpan generation, the constructed training sets were handed over to model training, which was accomplished by the program EDeN (Appendix A.2.4). EDeN accepts the training sets in gSpan format, extracts feature vectors and trains an SVM classification model based on the features found in the negative and positive sets (see Chapter 3). The computational pipeline which wraps up all the mentioned steps of set generation is described in Section 4.5.

4.4.2. Negative interactions

Negative set generation was performed similarly to positive set generation (Section 4.4.1). The main differences in the case of negative interactions concerned seed search and selection during filtering. As there do not exist any validated negative interactions of miRNA-target interactions, seed search was limited on mRNA regions that do not overlap with any of the identified miRNA-target interactions (full and partial hits) in the

five high-throughput experiments (Section 4.1). Thus, by using `intersectBed` (Appendix A.1.3), if an initial seed hit had an overlap with CLIP or CLASH sites, it was discarded. Additionally, negative seed search was only conducted on transcripts that contain positive target interactions of the respective miRNA that the seed scan was performed on.

Concerning the differences in selection, filtering was first conducted for positive interactions, which resulted in a reduced number of positive interactions. This filtering (described in 4.4.1) was also conducted for negative interactions, with the exception of site quality and energy filtering. In order to select negative interactions of the same size and hybrid characteristics, interactions were categorized into seed types. A seed type was defined as the number of base pairs plus the number of G:U base pairs in the seed. For example, if filtering of positive interactions for one miRNA resulted in 100 interactions of seed type "6-2" (6 seed base pairs, 2 G:U base pairs), 100 negative interactions with the same seed type were randomly chosen out of the usually much bigger number of negative interactions containing the sought seed type. This way, for every miRNA inside a dataset, two sets with the same size and similar hybrid characteristics were constructed.

4.5. The computational pipeline

This section describes the various steps taken by the computational pipeline, starting from negative set generation, filtering, gSpan generation and finally model training. In order to start the various scripts, positive dataset processing needs to be at the point where their hybrid statistics tables are present in the data directories. Additionally, tables containing site quality scores, mRNA sequences, protein crosslink and target site coordinates have to be available for the respective dataset to work with.

The concept of the pipeline is to create a training dataset (containing positive and negative interactions) for one specific miRNA belonging to one specific dataset, concentrating on one specific target region (5'UTR, CDS or 3'UTR). Computation for each script can be done either locally or on the workgroup cluster, using array jobs to run the same script in parallel with different settings. There are 3 main scripts representing the 3 main parts of the pipeline, which start additional scripts in order to complete their tasks. Those 3 parts will be described in the following subsections. A detailed pipeline description concerning script usage is given in Appendix A.3.

Part 1: Generation of negative sets

The first part begins with creating a list of mRNA IDs which are utilized for negative seed scanning. The script (`01-generate-negative-sets.pl`) expects a dataset ID, a target region and a miRNA ID (dataset-region-miRNA combination). Based on this information, the set of compiled positive interactions is searched for interactions with the given features, and their mRNA IDs are extracted. Alternatively, all the job can be

calculated on the cluster for an arbitrary list of dataset miRNAs, which is stored in a subdirectory table. In case of very few extracted mRNA IDs (e.g. if positive interactions of the given miRNA are rare), random mRNA IDs are added to the list such that the list contains a minimum of 30 mRNA IDs. This is important since seed type selection in the second part requires a sufficiently large amount of negative hits.

After mRNA ID list generation for each dataset miRNA, seed scanning starts on the listed mRNAs (see Section 4.2.2). Occuring hits are filtered depending on the given transcript region and overlaps with positive interaction sites (see Section 4.4.2). The remaining negative seed hits then are handed over to IntaRNA for hybrid prediction. Hybrid statistics are extracted as described (Section 4.2.3) and stored for subsequent filtering and gSpan generation.

Part 2: Set filtering and gSpan generation

The second part of the pipeline continues with set filtering and gSpan file generation. The script (`02-filter-sets-and-gspan.pl`) expects the same input as the first script, with the addition of optional filter settings for site quality and hybrid free energy cutoff. Once more, the calculation can be done locally for one miRNA or for a specified list of dataset miRNAs in parallel on the cluster. First, the positive set gets filtered and seed type occurrences in the remaining set get stored. Next, the negative set gets filtered the same way the positive set gets (see Section 4.4.1), expect for energy and site quality. Then, the same numbers of seed types that occur in the filtered positive set are extracted from the negative set (see Section 4.4.2). This usually results in identical numbers of positive and negative instances, which can then be converted into gSpan format and subsequently utilized for model training in the third part. If the number of a certain seed type in the positive set is higher than the number in the negative set, all members of the seed type are taken in in the negative set. This can be the case for certain miRNAs that exhibit a high number of 8mer seed motifs in the target sites, while the negative interactions percentage-wise contain far more 7mer or 6mer seed sites.

Based on the two filtered sets for each miRNA, generation of the gSpan files is performed as described (see Section 4.4). At the end of part two, negative and positive gSpan files have been created for every miRNA in the list, belonging to a certain dataset with hits on a certain transcript region.

Part 3: gSpan filtering and model training

Part three concludes the computational pipeline by applying optional filtering to the gSpan files, followed by feature decomposition and model training accomplished by EDeN (see Appendix A.2.4). The script (`03-filter-gspan-and-run-eden.pl`) again takes the dataset ID and target region ID as input, together with the miRNA ID and the optional filter parameters used in part two in order to identify the constructed gSpan files. In contrast to the second script, cluster mode does not correspond to the calculation of a list of miRNAs. This time, the list includes the gSpan filter settings as well as EDeN

parameters. It is therefore possible to test an arbitrary number of distinct settings and parameters for one miRNA (miRNA-dataset-region combination) in one parallel cluster run. The script has various gSpan filter parameters implemented, from deleting specific gSpan section information to deleting whole subgraph sections prior to running EDeN. The complete list of implemented filter parameters is described in Appendix A.3. Regarding EDeN settings, the two parameters r and d (see Section 3.3) as well as the cross-validation iterations can be set. Finally, script two outputs the performance of a distinct miRNA-dataset-region combination with a specific set of chosen filter and EDeN parameters. Based on the output, models and their performances can be analysed and optimized.

Results and Discussion

The following chapter presents the results obtained in the course of this thesis, along with discussing them. It is divided into three major parts, starting with the results obtained from analysing the utilized datasets. Part two reports the evaluation of the described miRNA extensions based on their influence on model performance, followed by the parameter optimization procedure conducted to find the best working models. Finally, part three describes the performances of the selected models regarding leave-one-out cross validation as well as their application on an independent test dataset.

5.1. Data analysis

The first few sections of this chapter deal with analyses conducted on the datasets utilized for this thesis (described in 4.1). Understanding their distinctive characteristics was defined as the first major thesis objective (see Section 1.3). In the course of data analysis, additional observations were made which should help to improve future prediction studies. Section 5.1.1 focuses on different mapping approaches in order to learn more about the datasets. Section 5.1.2 exemplifies the importance of target and miRNA abundance in miRNA-target interactions, and section 5.1.3 reports some additional observations such as transcript region differences and site quality filtering effects.

5.1.1. mRNA occupancy mapping

Exon border mapping

Exon junction complexes (EJCs) are known to reside approximately 24 nt upstream to exon borders (exon junctions) [79], acting e.g. in the translational quality control mechanism nonsense-mediated mRNA decay (NMD). Since most of these borders are reported to be occupied by EJCs, it can be questioned whether EJCs sterically influence miRISC binding and functioning. In order to test this, AGO1-4 PAR-CLIP sites were mapped to exon borders, either with all 41 positions or just with T-C position 21 (see Figure 5.1). In order to be utilized for the analysis, both border exons needed to be at least double the size of the plus and minus offset. This way, effects of neighbouring borders should be minimized. For each offset position, the number of overlapping sites was calculated with `intersectBed` and notated as position occupancy.

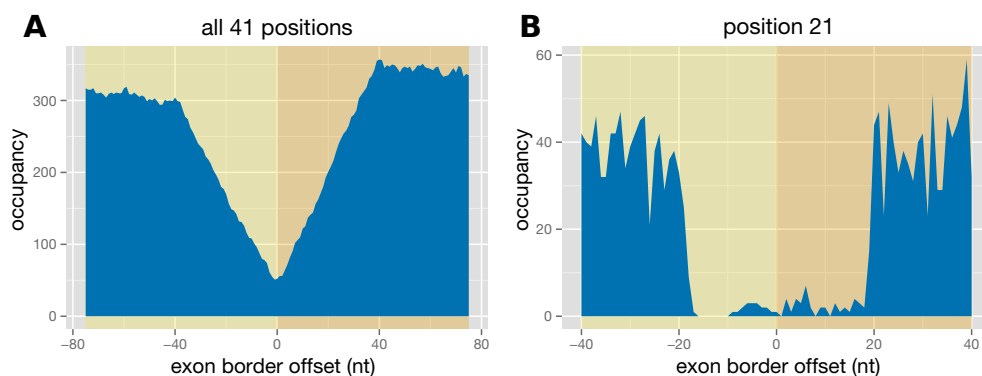


Figure 5.1.: AGO1-4 PAR-CLIP exon border occupancy. **A:** Full site mapping, encompassing all 41 site nucleotides. A minimum up- and downstream exon size of ≥ 150 nt resulted in 10337 analysed borders. **B:** Mapping of position 21 (T-C position). Minimum exon size ≥ 80 nt, resulting in 43805 analysed borders.

The two graphs show a clear cut in the size of the mapped sites, suggesting that AGO1-4 PAR-CLIP sites have been mapped to genomic sequences, which do not contain exon border overlaps. It has been stated in the AGO1-4 PAR-CLIP publication [10] that sequence reads were mapped to the human genome, human mRNAs and miRNA precursor regions. However, correspondance with the author confirmed the assumption, which also holds for the second AGO2 CLIP sites study [38]. This means that, by mapping sequence reads to the genome, target sites located at exon borders are discarded by these methods, which should be explicitly noted in future studies. Regarding EJC effects on CLIP site mRNA occupancy, the graphs do not seem to support this notion, with only slightly reduced downstream occupancy for the full site mapping, and no difference for position 21 mapping. Both assumptions were supported by mapping of CLASH sites (not shown), which due to reads mapped to mRNA sequences [39], did not

show the cut and only a slight downward motion on the left side for the full hit (taking the hybrid interaction regions) mapping.

Target site mapping

In order to get an impression of the transcriptome-wide mRNA occupancy of the five target site datasets, target sites were mapped to their respective dataset mRNA sequences (see Figure 5.2). For each mRNA region in the dataset, target site positions were mapped relatively to their location on the transcript region. For each location position, the number of mapped target sites was denoted as occupancy.

As we can see, the calculated differences in mRNA region occupancy (see Section 4.1) get reflected in the visualized mapping. All three AGO2 datasets (5.2 B-D) roughly have two thirds of their interaction sites located in the 3'UTR, while in the case of AGO1-4 PAR-CLIP mapping between CDS and 3'UTR is balanced (50,8 % CDS and 47,0 % 3'UTR). Other than that, all four CLIP data mappings adopt a similar shape, with both 3'UTR ends featuring more annotated target sites than the 3'UTR central region. This is in accordance to a miRNA-targeting principle observed by [48] (mentioned in 2.5), which states that target sites near the ends of long 3'UTRs are more effective than sites in the central part. Moreover, CLIP shapes seem to be independent of target site read coverage, since site quality filtering did not change the distinct shape at all (not shown).

In contrast to the CLIP mappings, AGO1 CLASH target sites show a clear preference to CDS over 3'UTR regions (61,5 % CDS and 34,1 % 3'UTR). The authors assumed that this might be due to read mapping to mRNA sequences rather than to genomic sequences. When reads mapped to exon borders were discarded, CDS percentage dropped to 50 %, while 3'UTR percentage increased to 42 %. Observations made in the previous section correspond to this discovery, which needs to be taken into account in future CLIP studies. Otherwise, these studies will completely ignore a seemingly considerable amount of target sites that span exon borders.

Taken together, although regional distribution of CLASH target sites seems to approximate the CLIP studies better by ignoring sites mapped to exon borders, there is still an obvious discrepancy concerning 3'UTR target numbers and 3'UTR inner regional distribution, especially between CLASH and the CLIP2 sets. Even though both methods used the same cell type and are technically related, CLIP and CLASH noticeably differ from each other regarding their mRNA occupancy profiles. The two CLIP studies show a higher degree of similarity, but still differences arise in the regional distribution, which might be due to variances in the utilized protocols. Further observed variations between the datasets will be discussed in the course of this chapter.

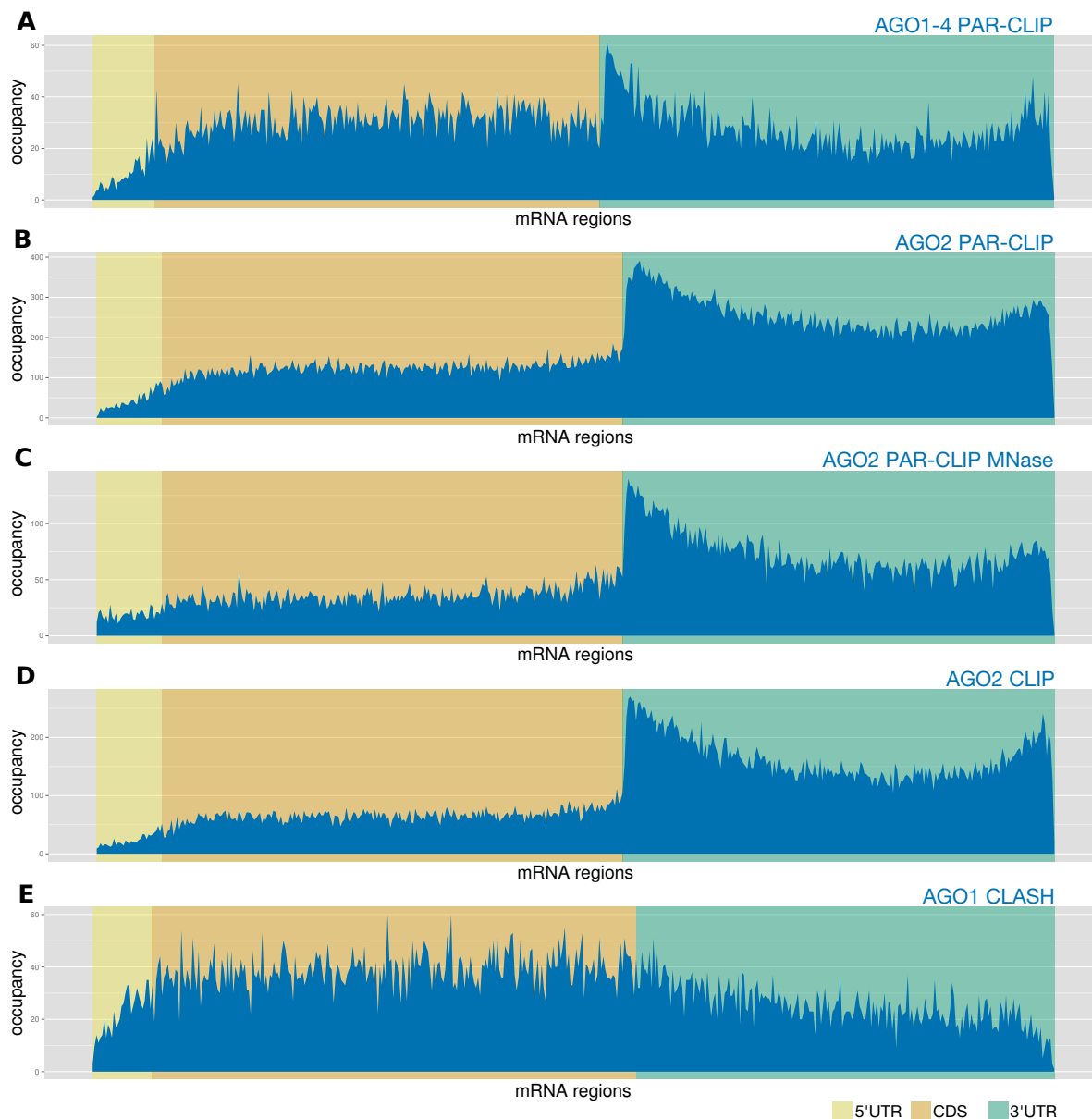


Figure 5.2.: mRNA occupancy profiles of the five target site datasets. For each dataset, single target site positions were mapped to mRNA sequences (AGO1-4 PAR-CLIP: position 21, AGO2: position 20, AGO1 CLASH: seed position 6). Each mRNA region is displayed with its average length in the respective dataset. Target site positions were mapped relatively to their positions in the target regions. The number of target site positions mapped to a certain region part denotes its occupancy. **A:** AGO1-4 PARCLIP (14317 site positions, 5733 mRNAs). **B:** AGO2 PAR-CLIP (90397 site positions, 9282 mRNAs). **C:** AGO2 PAR-CLIP MNase (44102 site positions, 8291 mRNAs). **D:** AGO2 CLIP (54367 site positions, 8494 mRNAs). **E:** AGO1 CLASH (16225 site positions, 6850 mRNAs).

Protein crosslink mapping

RNA-binding protein positions were mapped to crosslink-containing mRNA sequences analogous to target site mapping described in the previous section. Figure 5.3 shows the transcriptome-wide T-C crosslink positions (described in Section 4.1.5) distributed across the three transcript regions.

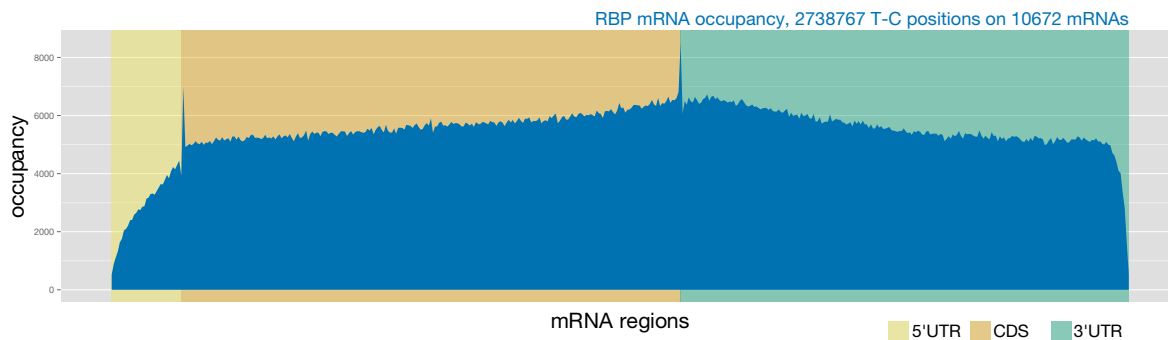


Figure 5.3.: mRNA occupancy profile of RBP T-C positions. 2738767 crosslink positions were mapped to 10672 distinct mRNAs that occur in the five target site datasets.

In contrast to AGO target site mapping, RBP T-C position mapping shows a more evenly distributed occupancy across the CDS and 3'UTR region. This is because we are not looking at the binding profile of a single RBP, but instead at the merged profile of 797 distinct RBPs [34]. Notably, since the graph actually maps T positions across an mRNA frame, the two spikes at the end of the CDS region are likely to depict the two start and stop codon thymines (or more precisely the uridines if the RNA sequence is considered), which are fixed in the mapping.

5.1.2. Importance of target and miRNA abundance

Mapping transcriptome-wide miRNA target sites to mRNAs allows one to take a look at frequently targeted transcripts. In this respect, it is interesting to see that many members of the miRNA machinery harbor multiple target sites on their own transcripts (e.g. AGO1-4, TNRC6A-C or DICER1). Table 5.1 shows the numbers of target sites mapped to the four AGO protein transcripts for all five target site datasets. At first sight, AGO1-4 PAR-CLIP and AGO1 CLASH contain more hits than the AGO2 sets, especially when incorporating the total number of target sites. Most notably, 88 target sites were mapped to the AGO1 transcript in AGO1 CLASH, making the transcript by far the most occupied mRNA in the dataset.

Investigating the causes, both CLASH and the original PAR-CLIP protocol (used for AGO1-4 PAR-CLIP) utilize HEK293 cells that stably express tagged AGO proteins in order to facilitate purification. AGO expression was induced by doxycycline 36 hours

Table 5.1.: Target sites on AGO transcripts. The number of target sites mapped on AGO transcripts for all five target site datasets. For the AGO2 datasets, site enrichment filtering with a cutoff ≤ 5 was applied (similar to AGO1-4 PAR-CLIP) in order to make the five sets more comparable. Like stated, the longest transcript variant was selected in case of multiple full hits. The four AGO transcripts were: NM_012199.2 (AGO1), NM_012154.3 (AGO2), NM_024852.3 (AGO3), NM_017629.3 (AGO4).

Dataset	Target sites	AGO1	AGO2	AGO3	AGO4
AGO1-4 PAR-CLIP	14317	24	8	26	3
AGO1 CLASH	17938	88	8	-	3
AGO2 CLIP	29471	18	13	3	8
AGO2 PAR-CLIP	25920	9	9	1	5
AGO2 PAR-CLIP MNase	26484	13	13	1	7

prior to UV radiation [39], which presumably resulted in overexpression and thus high mRNA abundance during UV radiation. It is reasonable to believe that both highly expressed mRNAs and miRNAs are more likely to participate in miRNA target interactions, which explains the high number of AGO1 target sites in the CLASH datasets. Indeed, recent kinetic analyses of AGO-miRNA mRNA targeting showed that miRNA function is determined by miRNA and target abundance [80]. Moreover, [81] reports that only the most abundant miRNAs exert target repression, while over 60 % of detected miRNAs had no measurable effect at all, indicating that the set of functional cellular miRNAs is much smaller than the profiled set.

As mentioned in Section 1.2.2, non-physiological concentrations can lead to non-physiological interactions, which constitutes a general problem of overexpression studies. While this might be the case for some of the detected AGO sites here, it is nevertheless interesting to look at these target sites in more detail. Theoretically, due to high target abundance, less stable hybrid interactions should occur more likely on these transcripts. Of the 88 AGO1 CLASH target sites, 68 resulted in a predicted IntaRNA hybrid. Among these, 50 % feature a noncanonical seed interaction (any 6mer seed), while for the whole AGO1 CLASH dataset (14101 predicted hybrids), this number was 43 %. Also, the percentage of non-contiguous G:U containing sites was 50 % versus 36 % in the whole set. Moreover, average hybrid energy of the 68 sites was -17.48 kcal/mol versus -19.53 kcal/mol in the whole set, indicating an increased amount of less-stable interactions, which supports the assumption that target abundance correlates with interaction stability.

Concerning the miRNAs that bind the 68 sites, there are 44 distinct miRNAs which on average bind to 146 sites in the CLASH dataset. This displays a huge difference compared to the average number of binding sites for all miRNAs in the dataset, which is 39. According to these results, high abundance targets may preferentially sequester highly abundant miRNAs. This is in agreement with an analysis of the AGO1-4 PAR-CLIP dataset [13], which showed that low-expressed miRNAs mainly bind to canonical target sites, while highly-expressed miRNAs also bind large amounts of noncanonical

sites. In conclusion, future target prediction studies could be optimized by incorporating differential prediction strategies depending on miRNA and target expression levels.

5.1.3. Further dataset characteristics

Overlap between datasets

In order to further examine the characteristics of the applied target site datasets, their site overlaps were calculated (see Table 5.2). Independent of the huge differences in target site numbers, discrepancies between the CLASH and CLIP sets again shows up in this comparison. Regarding CLIP and CLASH overlaps, the percentage of CLASH sites in CLIP sets and CLIP sites in the CLASH set is both noticeably smaller than the percentages among the CLIP sets. Even in the case of AGO1-4 PAR-CLIP, which features a more comparable distribution and number of target sites, there is less than 10 % overlap with CLASH target sites. In case of the percentage of CLASH sites overlapping with the CLIP sets (Figure 5.2 row 2), the highest amount of overlaps is obtained in the AGO2 PAR-CLIP column (30 %), which is by no means near the overlap e.g. of the AGO1-4 PAR-CLIP set (70.7 %). The result is even more surprising considering the fact that all five experiments were carried out in the same cell type (HEK293).

Table 5.2.: Overlapping target sites between the five datasets. The percentage of overlapping sites is given for each of the five target site datasets together with the number of target sites utilized in the calculation. Overlaps (defined as ≥ 1 common nt position between two sites) for each dataset on the left side with each of the datasets denoted as numbers on top are depicted in the table.

Dataset	Target sites	ID	1	2	3	4	5
AGO1-4 PAR-CLIP	14298	1	-	8.3	49.7	70.7	37.1
AGO1 CLASH	17938	2	7.9	-	19.6	30.0	22.5
AGO2 CLIP	54386	3	13.8	6.0	-	67.3	30.1
AGO2 PAR-CLIP	90417	4	12.3	5.5	42.0	-	27.7
AGO2 PAR-CLIP MNase	44109	5	13.0	8.6	36.9	53.7	-

Among the CLIP sets, further observations can be made. Particularly, AGO2 PAR-CLIP MNase seems to have less overlap than expected when considering the other two AGO2 sets, which both share more target sites with each other than with the MNase set. For example, AGO2 CLIP shares 67.3 % of its sites with AGO2 PAR-CLIP, while AGO2 PAR-CLIP MNase (which has even less sites) only shares 53.7 %. This observation has also been discussed by the authors [38], who concluded that the MNase dataset discovers a more unique set of target sites due to differential RNase treatment. In conventional PAR-CLIP (Section 1.2.3), crosslinked RNA segments are trimmed with RNase T1, which preferentially cleaves after G nucleotides, causing a bias towards recovering less

G containing target sites. Therefore, the authors supplemented RNase T1 with MNase (micrococcal nuclease), which preferably cleaves at A and T nucleotides, and also varied digestion conditions.

Comparison of predicted interaction energies

To further investigate the described bias, target site hybrid energies predicted by IntaRNA were compared across the five datasets (see Figure 5.4). Notably, both CLASH and MNase target sites exhibit better average interaction energies than sites from the other three CLIP sets. Moreover, CDS target sites feature better energies for all five datasets. This finding corresponds to the described observation (Section 2.5) that, in case of AGO1-4 PAR-CLIP, site complementarity in the CDS seems to be more strict than in the 3'UTR [14]. The high ranking of CLASH is somehow unexpected, giving the fact that there was no selection for the best target site energy as done with the CLIP sets. Ranking by GC content of the interaction sites however confirms this observation, with CLASH sites featuring the highest GC content (55.8 %), followed by the MNase set (46.7 %) and the remaining CLIP sets (AGO1-4 PAR-CLIP: 41.7 %, AGO2 PAR-CLIP: 40.6 %, AGO2 CLIP: 39.0 %). The better hybrid energies obtained for the two AGO2 sets compared to AGO1-4 PAR-CLIP likely originate from allowing IntaRNA to hybridize the seed along the whole 40 nt target site segment, while for AGO1-4 PAR-CLIP this target seed region was restricted (as described in 4.4.1).

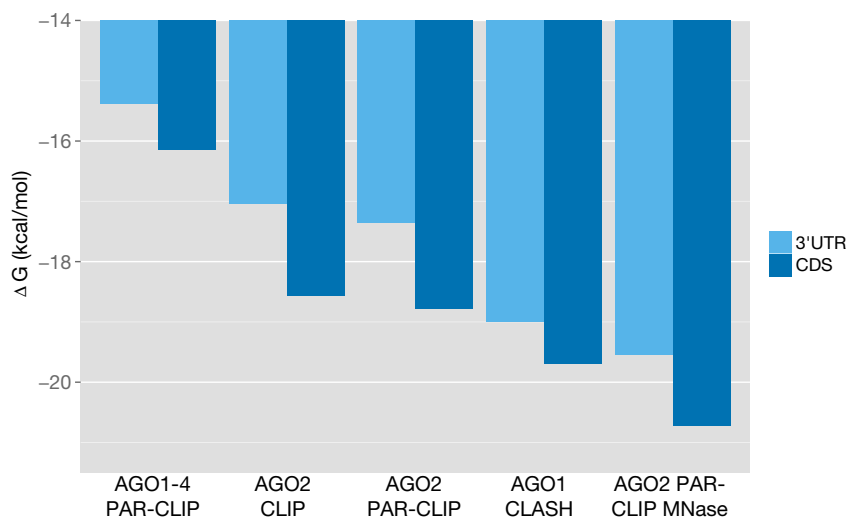


Figure 5.4.: Average hybrid energies of CDS and 3'UTR interactions. Average hybrid energies were calculated for all five datasets, both for the CDS and 3'UTR region. In case of CLIP, the average energy of the dataset was calculated based on the best IntaRNA energy for every target site (by pairing the sites with all top miRNAs and choosing the best). For CLASH, the IntaRNA energy of the identified miRNA-target pair was taken.

Regarding RNase exchange, G content was similarly distributed (AGO1-4 PAR-CLIP: 19.1 %, AGO2 CLIP: 19.6 %, AGO2 PAR-CLIP: 20.6 %, MNase set: 25.3 %, AGO1 CLASH: 30.1 %). The clear difference in G content between the MNase set and the

other two AGO2 sets thus confirms the effect of differential RNase treatment to recover target sites with distinct nucleotide compositions. Moreover, one could ask why CLASH exceeds the other sets in the GC content of their sites. As we have seen in former sections, differences exist both in the mapping profile and for target site overlap, from which it can be concluded that CLASH identifies sites with different characteristics. One of these characteristics might be the GC content, or related features such as secondary structure or site accessibility. Apparently, this can become an issue when utilizing CLASH as a test set for models trained on CLIP, and will thus be further discussed in Section 5.3.3.

Seed type occurrences in CLIP sets

As described in the last section, the MNase set differs from the other datasets in that it utilizes MNase instead of RNase T1, thus omitting the bias introduced for RNase T1 treated samples. In this regard, it is interesting to look at differences in seed type occurrences in the four CLIP datasets, which might arise from the different protocols. As stated by the authors [38], AGO2 PAR-CLIP, especially in combination with MNase treatment, yielded more miRNA seed-complementary sites than AGO2 CLIP. Furthermore, their amount was reported to be increased in highly enriched target sites. In this thesis, we wanted to see whether these observations also apply to hybrid interactions identified by IntaRNA in the CLIP datasets (see Figure 5.5). While increasing the applied site quality filtering, percentages of contiguous Watson-Crick seed-containing hybrids (6mer, 7mer, 8mer) were noted for each of the CLIP datasets. In case of AGO1-4 PAR-CLIP, site quality was defined as read coverage, while for the AGO2 sets, site enrichment was used (see Section 4.4.1).

Even though we did not use a direct seed scanning approach here, it is obvious that site quality (both enrichment for AGO2 and read coverage for AGO1-4) correlates with the percentage of contiguous Watson-Crick seed containing hybrids. This is an important finding, since site quality filtering was implemented and utilized in this thesis too. Filtering seems to be especially effective for enrichment of strong seed types, since 8mers show by far the steepest increase, outnumbering both 7mer and 6mer seeds in the high coverage target sites. This in turn implies that strong coverage correlates with strong seed pairing, which means that the CLIP protocols favorably detect strong seed type interactions.

Comparing the percentages among the CLIP sets, we can see that the MNase set indeed shows the highest amount of 8mer, but also 7mer sites, both encompassing nearly 20 % of all present seed types (as defined in Figure 5.5) in the 1 % most highly enriched sites. Also, percentages are higher in case of the AGO2 PAR-CLIP methods compared to AGO2 CLIP. Concerning AGO1-4 PAR-CLIP, percentages are lower for IntaRNA predicted hybrids, although this might again be due to the defined AGO1-4 PAR-CLIP seed region restriction mentioned in the last section.

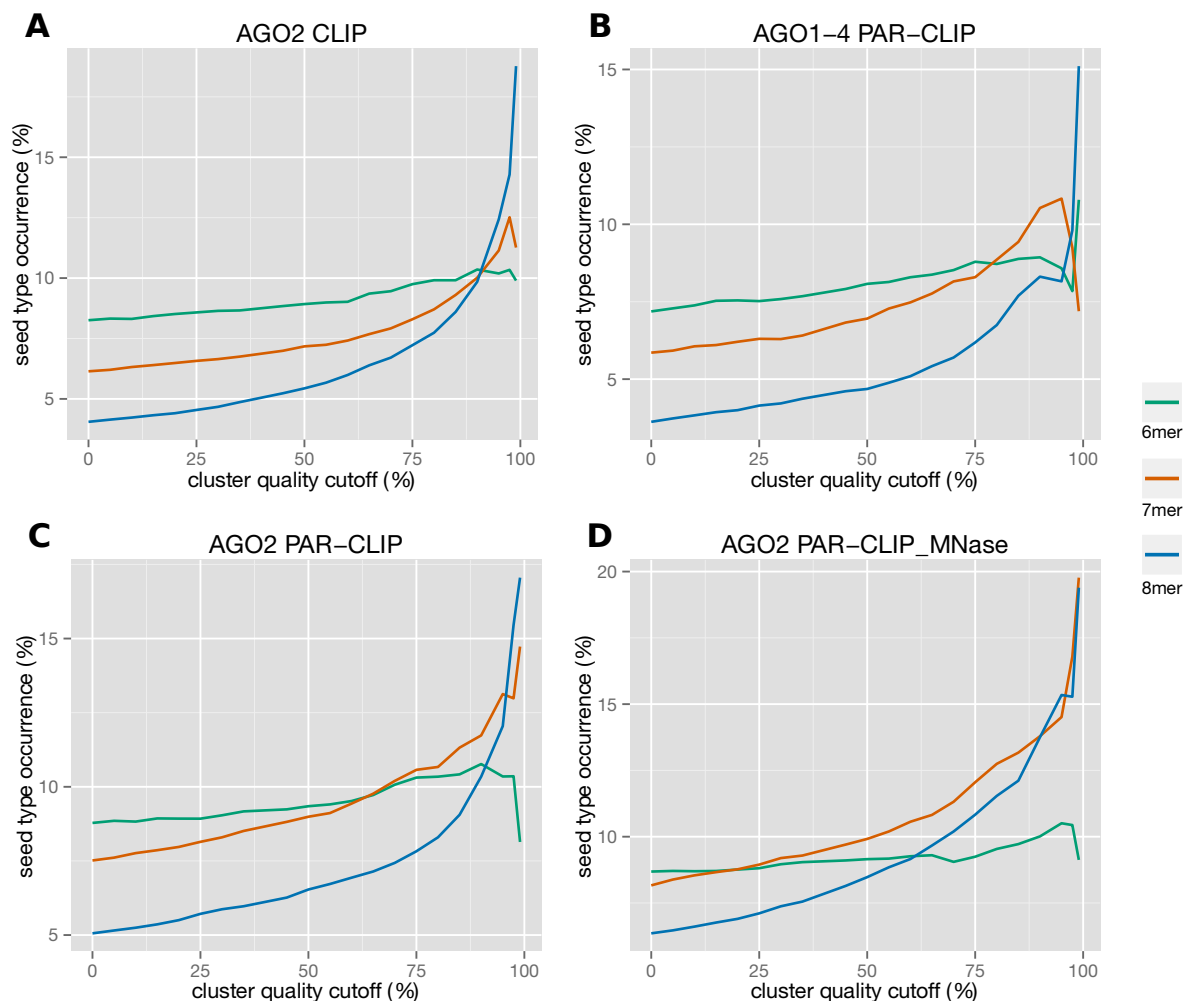


Figure 5.5.: Correlation of contiguous seed types with site quality. For each CLIP dataset, IntaRNA hybrid interactions were gradually (from 0 to 99 %) filtered by site (cluster) quality, each time noting the percentage of contiguous 6mer, 7mer and 8mer Watson-Crick seed types. Percentages were calculated based on dividing the seed types into four groups: Watson-Crick contiguous, Watson-Crick non-contiguous, contiguous G:U containing, non-contiguous G:U containing. Best CLIP IntaRNA hybrids were determined as described in Figure 5.4.

5.2. Model selection

Based on the introduced graph extensions described in Section 4.3.2, different extensions had to be tested concerning their effects on predictive performance. The AGO1-4 PAR-CLIP dataset was utilized for measuring these performances, since it features a sufficiently big set of predicted interactions (> 500 positive and 500 negative instances for the highly expressed miRNAs), which can be computed fast enough in case of many repetitions. Unless otherwise stated, all performances were measured using EDeN (see Section 4.5) with parameters $r = 2$, $d = 5$, as well as 10-fold cross validation. Also,

solely 3'UTR interactions were analysed. Section 5.2.1 examines the effects of the different interaction sections chosen for testing. Section 5.2.2 takes a look at the viewpoints extension (described in 3.2), and Section 5.2.3 reports the results of the graph kernel r and d parameter optimization for eleven different miRNA models.

5.2.1. Interaction sections

Testing of interaction sections encompassed evaluating the performances of five differently labeled interaction sections (see Section 4.3.2). All these sections were identical in that they contain the target-miRNA hybrid and span only the interaction region (defined in 4.3.1). The difference between the five sections resided in the different labeling of their miRNA and mRNA vertices. Therefore, depending on the labeling, different extents of information were included in the sections. Figure 4.2 shows the performance results of all five interaction sections. Each of the four training datasets was taken from the AGO1-4 PAR-CLIP set, comprising interactions of three single miRNAs as well as merged interactions of the ten top expressed miRNAs in the dataset.

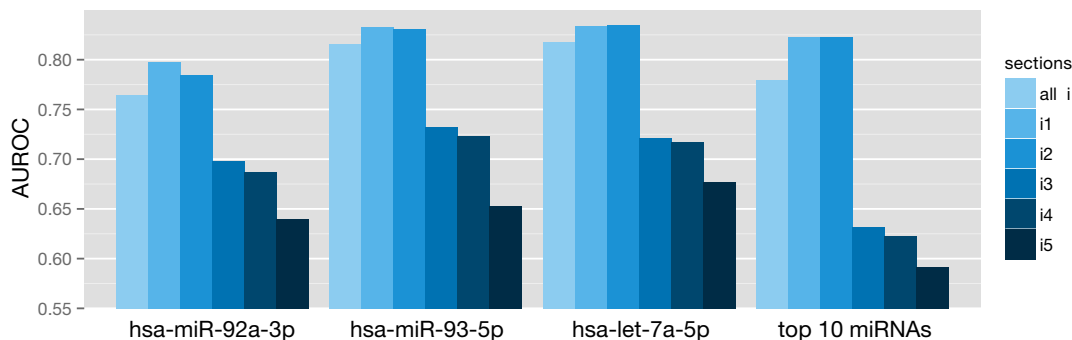


Figure 5.6.: Interaction section performance comparison. For each miRNA interaction set, performances of the five defined interaction sections as well as all interaction sections together are given. Added to the individual miRNAs is a merged interaction set of the ten top expressed AGO1-4 PAR-CLIP miRNAs. Performance is given in AUROC.

We can see that two of the interaction sections virtually perform equally well (i1 and i2), with good AUROCs mostly over 0.8. These interactions both contain the mRNA nucleotide information in their labels, as well as the miRNA nucleotide and position information (i1), or just the miRNA nucleotide information (i2). All other sections that use different or less informative labels perform considerably worse, which also explains the less optimal performance of all joined interaction sections. Comparing the four datasets, the merged miRNA dataset (top 10 miRNAs) performs equally as good as the two better individual miRNA sets, which indicates that certain characteristics of the interaction might be applicable for prediction independent of miRNA identity. Since the two sections i1 and i2 performed similarly good, section i1 was chosen to represent the interaction section information in the sequence and structure subgraph sections in the following experiments.

5.2.2. Viewpoints extension

Regarding the viewpoints concept (described in 3.2), performances were measured with or without set viewpoints as well as with different extensions. Figure 5.7 visualizes the results of the viewpoint measurements for four miRNA sets, on the sequence and structure model (described in 4.3). In short, the sequence model only contains sequence information, as well as (optionally) the hybrid information selected in the previous section. The structure model contains the sequence, hybrid interaction as well as the shrep secondary structure. For each interaction, three structure (shrep) sections exist.

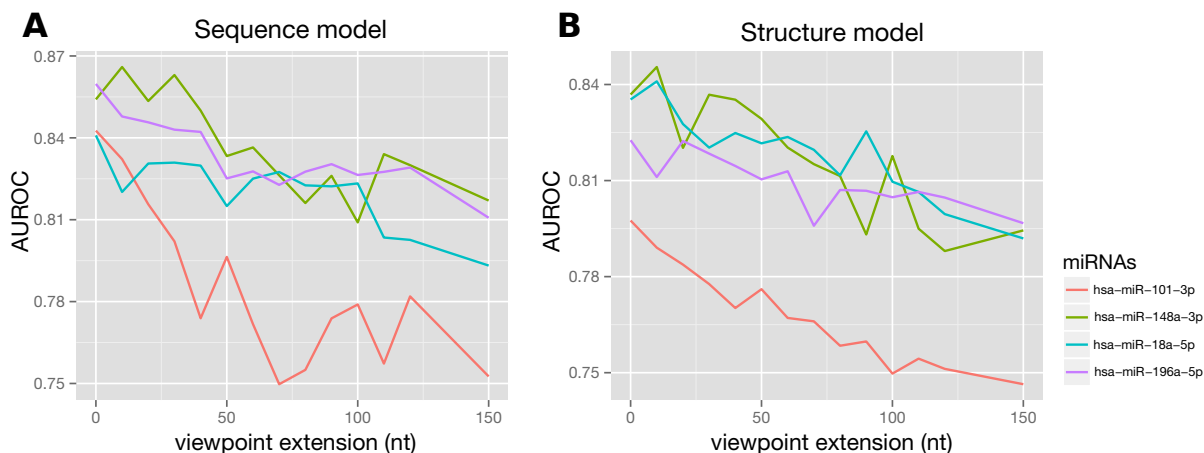


Figure 5.7.: Correlation between predictive performance and viewpoint extension.

The effect of viewpoint extension on both sides of the interaction region was observed measuring the performances of four miRNA sets in the sequence and structure model. The zero-value on the x-axis annotates the standard viewpoint, incorporating the interaction region. Value 150 annotates the full viewpoint, which is identical to the disabled viewpoint setting and was also measured this way. The stepsize of the measurement was 10, up to 120 nt.

It can be seen that for both models and all four miRNAs, AUROC performances decrease during viewpoint extension. An x-axis value of zero denotes the standard viewpoint (defined by the interaction region), which many times shows the best performance. In some cases, initially applied extensions perform slightly better, however this improvement is only small and inconsistent among the miRNAs. Note that the x-axis values for 150 nt were measured by disabling the viewpoints option. Since the extracted target segments were usually ~ 300 nt long, disabling viewpoints has the same effect as an extension of 150 nt. In this regard, setting no viewpoint always performs worse than keeping the originally set interaction region viewpoint. In an additional experiment (not shown), it was also tested whether reducing the viewpoint to the seed region (nucleotides 1-8) increases performance. This was not the case, since performance measures began to decrease or did not change for all tested miRNA sets. Comparing the structure and sequence model, it can be noted that the sequence model consistently performs better. Regarding these results, the standard viewpoint, comprising the interaction region, was chosen as the viewpoint setting for subsequent parameter optimization.

5.2.3. Graph kernel parameter optimization

Reducing set sizes

As described in Section 3.3, feature decomposition conducted by the graph kernel (EDeN) is controlled by two parameters (r and d), which needed to be optimized for all relevant miRNA models, in order to obtain optimal model performance. Prior to the optimization step, the observed influence of site quality on predictive performance was measured for five miRNA sets on the sequence and structure model (see Figure 5.8). Since the optimization procedure included a large amount of measurements, reducing the set sizes without losing too many instances and predictive performance to save time was a desired goal.

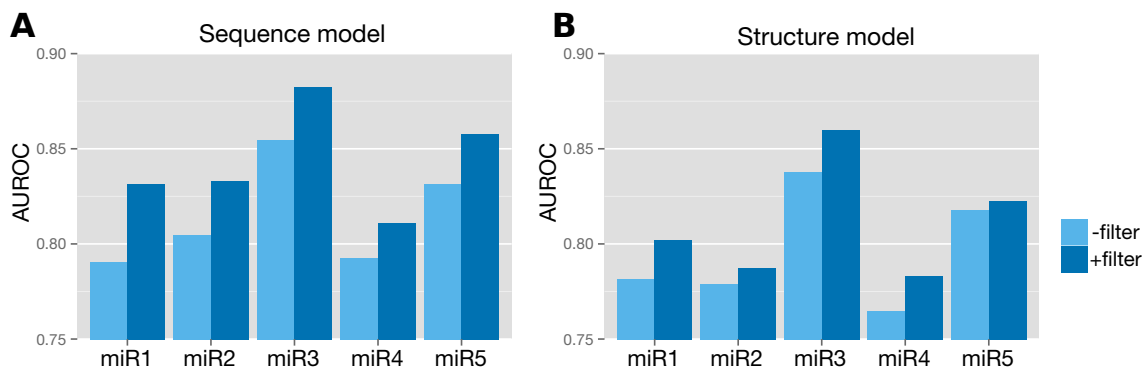


Figure 5.8.: Site quality filtering and its effect on performance. AUROC performance of the five top-expressed AGO1-4 PAR-CLIP miRNAs is given, measured with 50 % of sites filtered out by site quality (dark blue) and no site filtering (light blue). miR1: hsa-miR-19b-3p, miR2: hsa-miR-92a-3p, miR3: hsa-miR-93-5p, miR4: hsa-miR-103a-3p, miR5: hsa-let-7a-5p.

As we can see, applying site quality filtering by taking only the best 50 % improves prediction among all the observed miRNAs and models. This could be associated with the described increase in canonical seed interactions for the positive hybrids (Figure 5.5), although in case of 50 % the increase was still subtle for AGO1-4 PAR-CLIP. Importantly, the set numbers for the top expressed miRNAs were still above 500 negative and positive instances, which was defined as the minimum size to be utilized in this thesis. As with the measurement in the last section, the structure model performed worse than the sequence model.

Parameter optimization

The set of constructed miRNA models comprised eleven distinct models, which resulted from the combination of different informative features. Beside the hybrid, all model types either feature the sequence or the structure model, combined with additional prediction features. These additional features (described in Section 4.3) can be hybrid,


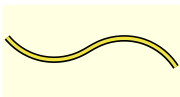
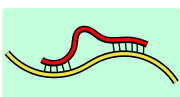
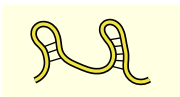
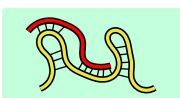

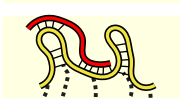
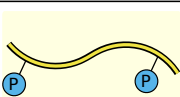
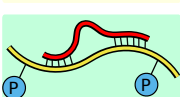
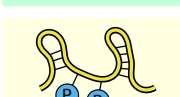
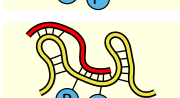
abstract structure or protein crosslink information. Table 5.3 depicts the eleven models together with the AUROC of the best performing r - d combination for five different highly expressed AGO1-4 PAR-CLIP miRNA sets. Importantly, these miRNAs were solely chosen for parameter optimization, and omitted in the following model training. For each miRNA-model combination, r values from 1 to 4 and d values from 0 to 6 were tested, resulting in 28 measurements for one combination and 1540 for all the possible combinations.

As a result, the sequence model containing the hybrid information (model 3), which was already found to perform good in previous sections, shows the best predictive performance among the models without additional protein information. Notably, the more sophisticated structure models perform worse than the sequence models, especially when abstract structure information is added. This seems to be a surprising fact at first sight, since the structure & abstract structure model (model 6) resembles the GraphProt model [73], which yielded good predictive performance results for AGO1-4 PAR-CLIP in the publication. However, it has to be said that parameter optimization was performed more rigorously in the case of GraphProt, using additional parameters for optimization. The fact that the choice of set parameters considerably influences predictive performance thus complicates comparisons. Moreover, negative instances were chosen differently in the publication, which makes it essentially impossible to compare the two results.

Considering the models with added protein information, the sequence model containing the hybrid again exhibits the best performance (model 9), and also shows the best performance among all eleven models. It can be argued that by adding protein crosslink information to the graphs, positive instances might become recognizable due to added AGO crosslink information, which is of course also present in the protein crosslink set. While this cannot be checked directly (since the crosslinks do not contain labels), the authors of the protein study [34] noted that 76 % of analysed PAR-CLIP sites contained T-C changes that were present in the protein profile. Looking at the target sites of the five top expressed miRNAs in AGO1-4 PAR-CLIP, we observed a similar number (69.8 %), while for the corresponding negative interaction regions, only 49.1 % contained one or more T-C crosslink positions.

Concluding from these results, the three models highlighted in green (model 3, model 5, model 9) were chosen for subsequent model evaluation. Model 5 was additionally taken into account, since it comprises the best performing structure model. Otherwise, no structure information would have been present anymore in later comparisons. The optimal parameters found for the three models were $r = 3$ and $d = 6$ for model 3, $r = 2$ and $d = 6$ for model 5, and $r = 3$, $d = 5$ for model 9.

Table 5.3.: Parameter optimization results for 11 models and 5 miRNAs. The AUROC of the best performing r-d setting is given for each miRNA-model combination. For each model, the r-d combination that ranked best among all five tested miRNAs was selected as the optimal model setting and subsequently used in model evaluation. Tested r-values: 1 – 4, tested d-values: 0 – 6. Dataset: AGO1-4 PAR-CLIP. Settings: site quality best 50 %, 10-fold-cross-validation. miR1: hsa-miR-18a-5p, miR2: hsa-miR-26a-5p, miR3: hsa-miR-101-3p, miR4: hsa-miR-148a-3p, miR5: hsa-miR-196a-5p.

	Model	Description	miR1	miR2	miR3	miR4	miR5
1		hybrid	0.845	0.781	0.832	0.872	0.854
2		sequence	0.828	0.805	0.842	0.864	0.836
3		sequence & hybrid	0.866	0.813	0.853	0.872	0.866
4		structure	0.794	0.748	0.784	0.826	0.802
5		structure & hybrid	0.844	0.764	0.821	0.858	0.840
6		structure & abstract structure	0.706	0.669	0.695	0.734	0.744
7		structure, hybrid & abstract structure	0.800	0.701	0.766	0.801	0.810
8		sequence & protein profile	0.864	0.836	0.841	0.885	0.887
9		sequence, hybrid & protein profile	0.882	0.850	0.859	0.897	0.905
10		structure & protein profile	0.824	0.809	0.797	0.863	0.849
11		structure, hybrid & protein profile	0.883	0.825	0.843	0.881	0.887

5.3. Model evaluation

Based on the results in the previous sections, three models were chosen for subsequent model evaluation. All three models feature the interaction region chosen in Section 5.2.1, as well as the standard viewpoint comprising the interaction region (evaluated in Section 5.2.2). Section 5.3.1 reports performances on the merged CLIP dataset with the three chosen models, and compares it to AGO1-4 PAR-CLIP set performance. Section 5.3.2 shows the generalization ability of the models, by applying leave-one-out cross validation with individual miRNAs. Finally, Section 5.3.3 presents the results of testing the CLIP trained models on the CLASH dataset.

5.3.1. CLIP dataset performances

In order to utilize the remaining CLIP datasets, one big dataset comprising all for CLIP sets (AGO1-4 PAR-CLIP, AGO2 CLIP, AGO2 PAR-CLIP, AGO2 PAR-CLIP MNase) was constructed. Site quality was normalized for filtering as described in Section 4.4.1, and miRNA lists were merged, keeping only the ten most highly expressed miRNAs in both AGO1-4 and the AGO2 sets. In order to compare the results, AUROC performances of the three chosen models were first measured for the ten top expressed AGO1-4 PAR-CLIP miRNAs (Figure 5.9).

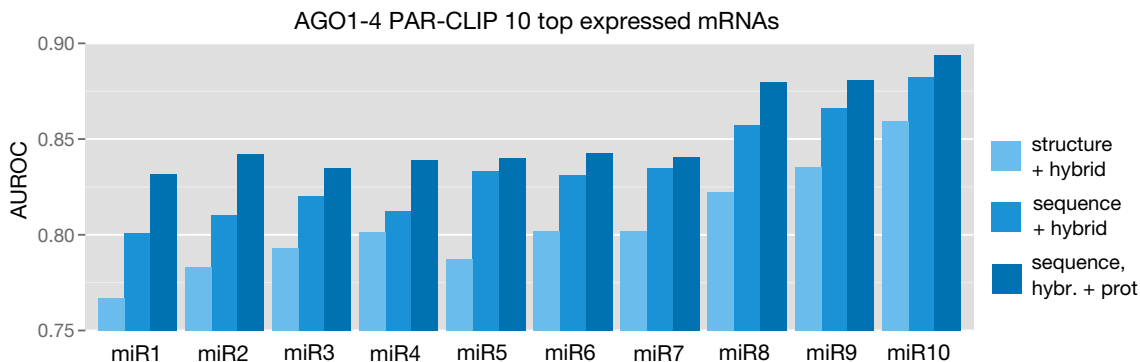


Figure 5.9.: Performances of 10 top expressed miRNAs in the AGO1-4 PAR-CLIP dataset. AUROC performance of the three chosen models (sequence + hybrid, structure + hybrid, sequence + hybrid + protein information). Each miRNA dataset was filtered by site quality (50 %), and contained at least 500 positive and 500 negative instances. Testes miRNAs: miR1: hsa-miR-30e-5p, miR2: hsa-miR-103a-3p , miR3: hsa-miR-21-5p, miR4: hsa-miR-423-3p, miR5: hsa-miR-92a-3p, miR6: hsa-miR-19b-3p, miR7: hsa-miR-10a-5p, miR8: hsa-let-7a-5p, miR9: hsa-miR-301a-3p, miR10: hsa-miR-93-5p.

In agreement with the previous-section results, the stucture model performs worst across the board, while all miRNA sets show a good predictive performance for the sequence + hybrid model. The model consistently performs with an AUROC of ≥ 0.8 , which

gets further topped by adding protein information (sequence, hybrid + protein model). However, these generally good performances did not sustain in the merged CLIP dataset (Figure 5.10). We can see that all measurements excluding one protein model result comprise ≤ 0.8 AUROC. Although the relative ranking between the models stayed the same, their overall performances went considerably worse.

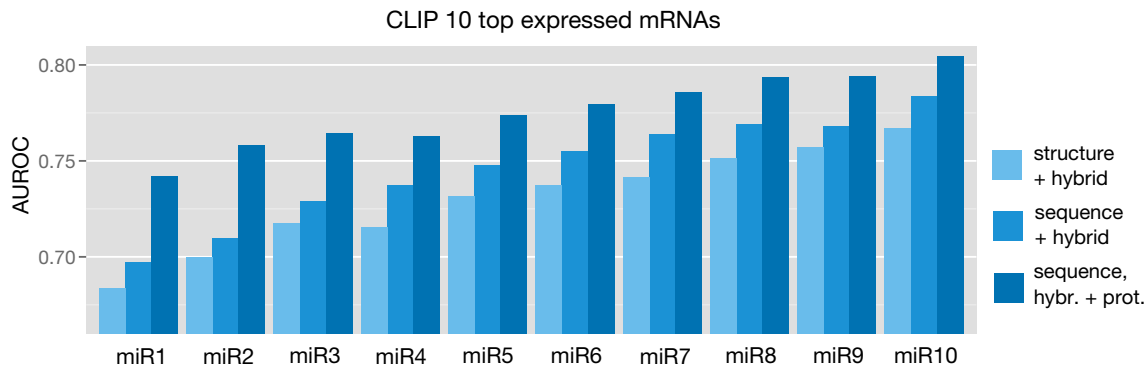


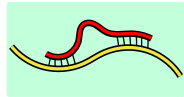
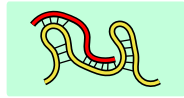
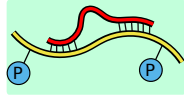
Figure 5.10.: Performances of 10 top expressed miRNAs in the merged CLIP dataset. AUROC performance of the three chosen models (sequence + hybrid, structure + hybrid, sequence + hybrid + protein information). Each miRNA dataset was filtered by site quality (50 %), resulting in instances per miRNA (+ and -) from 8634 to 31891. Testes miRNAs: miR1: hsa-miR-21-5p, miR2: hsa-miR-10a-5p, miR3: hsa-miR-92a-3p, miR4: hsa-miR-30e-5p, miR5: hsa-miR-93-5p, miR6: hsa-miR-16-5p, miR7: hsa-miR-103a-3p, miR8: hsa-miR-19b-3p, miR9: hsa-let-7a-5p, miR10: hsa-miR-301a-3p.

Searching for explanations, we have learned about the differences between the datasets in previous sections. This indeed could become a problem, especially since the MNase set which was also included in the set was shown to comprise distinctly featured target sites. During the construction of the dataset, these differences were not as apparent as in the actual training phase. It could therefore be appropriate to separately train and test the sets in future studies, before trying to merge them. Since the CLIP set performance turned out to be unsatisfactory, remaining experiments were conducted with the AGO1-4 PAR-CLIP dataset.

5.3.2. Assessing generalization ability

Up to this point, model performance was evaluated by using single miRNA sets in conjunction with 10-fold cross validation. In order to assess model performance regarding its ability to generalize beyond the instances in the training set, leave-one-out cross validation was performed on a set of ten top expressed miRNAs. Precisely, every miRNA set was utilized once as a test set, while the remaining 9 miRNAs were used for model training. Performance was then evaluated by taking the average measures out of the 10 test-training phases. Table 5.4 sums up the results for the leave-one-out cross validation.

Table 5.4.: Leave-one-out cross-validation on AGO1-4 PAR-CLIP. Leave-one-out cross-validation was performed for interactions of the ten top expressed AGO1-4 PAR-CLIP miRNAs. In each of the ten iterations, one distinct miRNA was used for testing the model trained on the 9 remaining miRNAs. Average quality measures were taken from the individual ten performance measurements.

	Model	Sensitivity	Specificity	Precision	AUROC
3		0.677	0.686	0.687	0.753
5		0.726	0.629	0.664	0.744
9		0.776	0.677	0.708	0.806

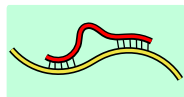
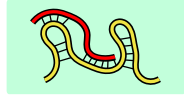
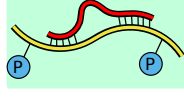
As we can see, the overall performance of the three models decreases as expected for this increasingly difficult prediction task. The performance ranking still holds up for the three models, with the protein model (model 9) being the best performing one, followed by the sequence model (model 3) and the structure model (model 5). In general however, the performance is still fair, which is also denoted by the fair sensitivity and specificity performance. Noticeably, the structure model is almost as good as the sequence model, comprising an even higher sensitivity.

5.3.3. Testing the trained models

Testing the predictive model on an independent test dataset comprised the last objective of this thesis. So far, three miRNA models had been chosen for this task and trained on AGO1-4 PAR-CLIP data, with fair performance results in leave-one-out cross validation. AGO1 CLASH was chosen as the independent test dataset, since it already had been analysed and converted into the desired data formats. For the actual test, the ten top expressed AGO1-4 PAR-CLIP miRNA sets, whose distinct cross validation performances had been measured in Sections 5.3.1 and Section 5.3.2, were merged into one big dataset, containing 7142 positive and 7121 negative interactions. Subsequently, a model was generated based on utilizing these instances as training sets (detailed in A.2.4). The generated model was then evaluated on the whole CLASH 3'UTR dataset, which contained 3907 positive and 3768 negative hybrid interactions. Table 5.5 shows the results of the test run.

At first sight, performances of the three models further drop in comparison to the obtained leave-one-out cross validation results (Table 5.4). The sequence protein model still performs best (model 9), while the structure model (model 5) works slightly better

Table 5.5.: Model performance on the CLASH test data. Performance of the three chosen models on the CLASH test dataset. The merged AGO1-4 PAR-CLIP dataset contained 7142 positive and 7121 negative instances, while the CLASH dataset comprised 3907 positive and 3768 negative instances.

	Model	Sensitivity	Specificity	Precision	AUROC
3		0.203	0.808	0.564	0.542
5		0.230	0.818	0.567	0.561
9		0.431	0.768	0.658	0.643

than the sequence-hybrid model (model 3) this time. Notably, sensitivity is relatively low, while specificity is substantially increased. This means that the model arguably has less problems with the correct classification of negative instances as with the correct classification of positive instances. This actually perfect sense, since the negative instances were chosen the same way for the CLIP and the CLASH sets. The low sensitivity thus points to distinctive target site features in the CLASH dataset, which seem to be unknown or not significantly enriched in the CLIP model.

Recalling the various described differences between CLASH and the CLIP datasets, the result may be less surprising than initially perceived. In general, miRNA target prediction performances are far from being satisfying (for details see Section 1.2.3). One reason for this shortcoming might simply be that training sets contain features which are not as important or even irrelevant in the test sets. For example, [82] showed that models trained on CLIP datasets generally perform good on other CLIP datasets, but poor on expression data and vice versa. Finally, in order to exclude the inability to learn any distinguishable features from the CLASH data set, the performance of both the CDS and the 3'UTR CLASH data as training sets was tested for the three models using 10-fold cross validation (Figure 5.11).

As we can see, predictive performance for both CDS and 3'UTR CLASH sets is comparable to the performance obtained from the CLIP 10-fold cross validation measurements. Interestingly, the models work quite well, considering that 366 distinct miRNAs contributed hybrid information to the models. Most surprisingly, the structure model performs the best both for the CDS and the 3'UTR hybrid interactions, while the protein model performs worst in the case of the CDS set. This result yet reflects another difference between the CLIP and CLASH datasets, which could explain the slightly better, second-place performance of the structure model in table 5.5. The increased GC content of CLASH target sites (noted in 5.1.3) might also be important in this context,

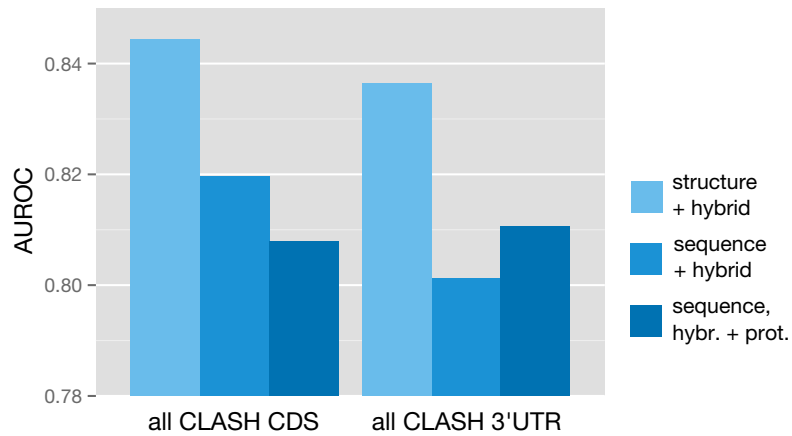


Figure 5.11.: Performance of all CLASH miRNA-target sites combined. AUROC performance of the three chosen models (sequence + hybrid, structure + hybrid, sequence + hybrid + protein information) was tested on CDS and 3'UTR region CLASH target sites. Number of instances for the two sets: CDS (6834+ 6737-), 3'UTR (3907+ 3768-). Number of miRNAs in the two sets: 366.

since GC rich regions preferably form secondary structures. In conclusion, intra-dataset predictive models worked better for both CLIP and the CLASH datasets, while their described distinctive features seem to be responsible for the less optimal results obtained in inter-dataset testing.

Conclusion and Outlook

In the course of this thesis, a novel graph-based machine learning model was extended in order to be utilized for miRNA target prediction. High-throughput datasets were compiled to train and test the generated models. A rich repertoire of non-canonical seed sites as well as canonical seed sites was successfully integrated into the predictive model. Various graph extensions were tested and incorporated as well. Optimally performing models were identified by graph kernel parameter optimization and subsequently trained and tested. Concerning the results, intra-dataset model training resulted in good predictive performances (AUROC > 0.8), while testing performed on an independent dataset was shown to still have room for improvement. In particular, uncovered differences in the utilized datasets seem to have an important impact on predictive performance and should thus be considered during subsequent model refinement.

Beside the discussed differences in dataset characteristics, further improvements can be made regarding several issues. First of all, parameter optimization could be conducted more thoroughly, including additional parameters that were e.g. used for the GraphProt model. Unfortunately, this was not possible given the limited amount of time for this work. Moreover, a more precise integration of miRNA and target abundance as described in Section 5.1.2 should lead to noticeable improvements in predictive performance. Also, the integration of increasingly available, more specific protein binding site information should lead to a more precise model, which was demonstrated in the case of Pumilio binding sites [83]. Another important factor seems to be the presence of multiple binding sites in close vicinity, as discussed in Section 2.5, which could be incorporated into the graph model as well.

Summing up, although the resulting predictive performance of the generated models still leaves a lot to be desired, the implemented graph-based approach nevertheless provides a flexible and easily extendable prediction environment, which allows the insertion of new graph features combined with rapid cluster evaluation thanks to the implemented computational pipeline.

APPENDIX A

Computational Details

Appendix A details various computational details in data pre-processing (A.1), training dataset generation and utilization of the data in model training and testing (A.2). Furthermore, a computational pipeline usage description is given in A.3.

A.1. Data pre-processing

A.1.1. Sequence feature extraction

In order to extract sequence features such as exon or CDS annotations from RefSeq genes, the BioPerl (version 1.6.901-2) NCBI GenBank database interface was utilized¹. The following Perl code exemplifies this:

```
#Include the GenBank perl module.
use Bio::DB::GenBank;
# Example refseq ID.
my $refseqID = "NM_012154.3";
# Create database object.
my $db = Bio::DB::GenBank->new;
my $seq = $db->get_Seq_by_acc($refseqID);
# Get transcript sequence.
```

¹<http://search.cpan.org/~cjfields/BioPerl-1.6.1/Bio/DB/GenBank.pm>


```
my $sequence = $seq->seq();
# Get CDS and exon end coordinates.
my $feat; my $start; my $stop; my $tag;
my @exonEnds; my $exonStop;
foreach $feat ( $seq->top_SeqFeatures() ) {
  foreach $tag ( $feat->primary_tag() ) {
    if ( $tag eq 'CDS' ) {
      $start = $feat->start;      # CDS start position
      $stop = $feat->end;        # CDS end position
    }
    if ( $tag eq 'exon' ) {
      $exonStop = $feat->end;    # Exon end position
      push(@exonEnds, $exonStop); # Save
    }
  }
}
```

A.1.2. Target sequence mapping

Alignment of the target site sequences was accomplished by constructing a local BLAST database with `formatdb`, containing the downloaded RefSeq hg19 transcript collection (described in 4.1.4). The `blastn` (Nucleotide-Nucleotide BLAST 2.2.25+) tool was then used to search the database for the target sequences in order to recover the target site positions and transcript IDs. First, `formatdb` is called with parameter `-p F` to build a nucleotide BLAST database based on the FASTA sequences in `refMrna-hg19.fa`:

```
formatdb -i refMrna-hg19.fa -p F
```

This generates three database files (`refMrna-hg19.fa.nhr`, `refMrna-hg19.fa.nin`, `refMrna-hg19.fa.nsq`) which can then be used in conjunction with `blastn` to search the database for an input sequence stored in `query.fa`:

```
blastn -query query.fa -db refMrna-hg19.fa -dust 'no' -num_threads 4 -strand plus
      -task blastn-short -evalue 1e-04 -outfmt "7 qseqid sseqid slen pident qstart qend
      sstart send qseq evalue"
```

By default, `blastn` uses the DUST filter for query sequences, which masks simple sequence repeats (low-complexity sequences). The option was turned off (`-dust 'no'`), since some full hits were ignored by the filter. Multithreading was used (`-num_threads 4`), as well as `-task blastn-short` which optimizes the algorithm for short query sequences. The e-value threshold was set to 0.0001 for the CLIP datasets, with query sequence lengths of 41 and 40 nt. In case of CLASH, the threshold was less strict with 0.001, since CLASH query sequences feature various lengths from 18 to 119 nt. The output format was specified with the `-outfmt` option, and the hit with the lowest e-value was taken. In case of multiple hits with the same e-value, the longest transcript was chosen.

A.1.3. BED format operations

The BED format was designed to encode genome browser annotation tracks¹ and can also be used to calculate overlaps between genomic or transcriptomic regions. The following three fields are mandatory in order to annotate a genomic or transcript region in BED format: the name of the chromosome or transcript, followed by the start and end coordinate, with tab-separated entries. Notably, the first coordinate needs to be zero-based for correct output, while the second is one-based, which frequently leads to confusions. As an example, the following three rows annotate the three transcript regions of an mRNA:

```
NM_001142640.1      0      600      5utr_NM_001142640.1      0      +
NM_001142640.1      600     5781     cds_NM_001142640.1      0      +
NM_001142640.1      5781    9794     3utr_NM_001142640.1      0      +
```

The software package BEDTools² was utilized to calculate overlaps between BED format files. In order to calculate overlapping BED rows between two files, `intersectBed` was used, which is illustrated by the following call:

```
intersectBed -a file1.bed -b file2.bed -u -f 1 > result.bed
```

This results in writing all the regions in `file1.bed` that fully overlap (`-f 1`) with regions in `file2.bed` to the `result.bed` output file. The `-u` option defines that each region is only reported once, even if there are several overlaps. Various other options can be set in order to achieve the desired result. Beside BEDTools, the BEDOPS³ package has also been used to calculate more complicated set operations (e.g. overlaps between more than two files).

A.1.4. Human genome assembly conversion

In order to locally convert genomic coordinates to a different genome assembly version, the `liftOver` executable plus the necessary conversion file (`.chain`) was downloaded from UCSC⁴. The tool then takes the genomic BED file which needs to be converted together with the `.chain` file, the output file and a log file for unmapped entries, and executes the conversion. This call exemplifies an hg18 to hg19 conversion:

```
./liftOver file-hg18.bed hg18ToHg19.over.chain file-hg19.bed unmapped-hg18
```

¹<http://www.genome.ucsc.edu/FAQ/FAQformat.html>

²<http://code.google.com/p/bedtools/>

³<https://bedops.readthedocs.org>

⁴liftOver: <http://hgdownload.cse.ucsc.edu/admin/exe/>

⁵Conversion file: <http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/>

A.2. Training data generation and utilization

A.2.1. Regular expression seed scanning

Perls' regular expressions were utilized for seed scanning mRNA sequences. Subroutines which accept the miRNA seed sequence and return the regular expression strings were implemented. The following Perl code illustrates the creation of three distinct regular expressions:

```
sub construct_some_regexes {
    # Seed sequence as input.
    my($seed) = @_;
    # Make reverse complement (seed motif).
    my $seedRC = reverse($seed);
    $seedRC =~ tr/ACGU/UGCA/;
    # Split seed characters into array.
    my @ch = split //, $seedRC;
    # Arbitrary nucleotide.
    my $insert = "[A|C|G|U]";
    my $pos;
    # Add regular expressions for G:U pairs.
    for (my $i = 0; $i < 8; $i++) {
        if ($ch[$i] eq "A") {
            $pos = "[A|G]";
        } elsif ($ch[$i] eq "C") {
            $pos = "[C|U]";
        } else {
            $pos = $ch[$i];
        }
        $ch[$i] = $pos;
    }
    # Construct a 2-7 match.
    my $regex1 = $insert.$ch[1].$ch[2].$ch[3].$ch[4].$ch[5].$ch[6].$insert;
    # Construct mRNA bulge with bulge between miRNA positions 5 and 6.
    my $regex2 = $insert.$ch[1].$ch[2].$insert.$ch[3].$ch[4].$ch[5].$ch[6].$insert;
    # Construct miRNA bulge with bulge between miRNA positions 5 and 6.
    my $regex3 = $ch[0].$ch[1].$ch[3].$ch[4].$ch[5].$ch[6].$insert;

    # Return regular expression search string.
    return "$regex1|$regex2|$regex3";
}
```

In case of a mRNA bulge, the search string is extended to 9 positions, while for miRNA bulges, the search string only contains 7 positions. Mismatches, the fourth distinct class of regular expressions used (beside classes that comprise contiguous or the 2 bulge matches), were constructed in a similar way. Here, depending on the nucleotide, a mismatch expression has to be inserted. For example, in the case of a G nucleotide,

which can only pair with C, "[C|A|U]" gets inserted into the search string. Notably, when using `$insert`, the 2-7 match in the example can also spot 7mer or 8mer matches.

The second step involves the actual search, using the created regular expression search string in conjunction with Perl's match operator:

```
# Get the regular expression search string.
my $regex = construct_some_regexes($seed);
# Save hits in BED format.
my $bedHits = "";
# Scan the sequence.
while ($sequence =~ /$regex/g) {
    # Get the sequence positions of the hit.
    $start = $-[0]; # zero-based.
    $end = $+[0];
    # Continue search at position pos($sequence).
    pos($seq) = $-[0] + $1 + 10;
    $bedHits = $bedHits . "$sequenceID\t$start\t$end\t$seedID\t0\t+\n";
}
```

If the search string is found, the positions of the hit get stored in a BED file along with the sequence ID and additional information such as a seed ID. In this example, the search is not continued right after the hit, but instead after an offset of its length plus 10 nucleotides. This was done for the negative seed scanning, in order to prevent huge numbers of reported seed hits.

A.2.2. IntaRNA hybrid prediction

In this thesis, a unofficial beta version of IntaRNA¹ (version 1.2.6) was utilized for computing the minimum free energy miRNA target hybrid. In addition to the official version, it allows the definition of a seed region on the target, which was used for the AGO1-4 PAR-CLIP dataset (see Section 4.2.3). This is an example call:

```
IntaRNA -t mrna.fa -m mirna.fa -o -p 6 -u 2 -a 0 -b 0 -f 1,8 -e 21,30
```

IntaRNA expects two input sequences (`-t` for target and `-m` for mirna) in FASTA format, followed by several optional options. Detailed output (`-o`), a specified minimum number of base pairs in the seed (`-p 6`), the maximum number of unpaired seed bases in both sequences (`-u 2`), disabled accessibility calculation for sequences a and b (`-a 0`, `-b 0`), and the defined seed regions on the miRNA (`-f 1,8`) and the target (`-e 21,30`) were usually set.

¹<http://rna.informatik.uni-freiburg.de:8080/IntaRNA/Input.jsp>

A.2.3. FASTA to gSpan conversion

In order to generate FASTA sequences into gSpan formatted graph files, the existing Perl script `fasta2shrep_gspan.pl` was utilized, which allows various settings in order to obtain the desired result. A typical example with settings used in this thesis:

```
perl fasta2shrep_gspan.pl -fasta segment.fa -M 3 -abstr -seq-graph-t -vp -i 100
```

File `segment.fa` contains the RNA sequence for shrep prediction and conversion into the gSpan format. The output includes the shreps of the three most probable shapes (`-M 3`), integrates abstract structure information (`-abstr`), adds an unstructured sequence subgraph (`-seq-graph-t`) and sets viewpoints (`-vp`). The area of set viewpoints is defined by the sequence in the FASTA file, where uppercase nucleotide letters mark the viewpoint region. The option `-i 100` enables structure sampling, which speeds up the shape calculation.

A.2.4. EDeN feature extraction and model training

EDeN (Explicit Decomposition with Neighborhoods)¹ combines NSPDK Kernel feature decomposition with subsequent machine learning model training and testing (see Sections 3.3 and 3.4). There are various parameters in order to specify the desired classification or regression task. The following EDeN call was used for feature extraction:

```
./EDeN -i inputData.gspan -a FEATURE -r 2 -d 5
```

The input data file contains the positive and negative interaction graphs in gSpan format, of which EDeN decomposes the features and stores them in a feature file. The two parameters that control feature decomposition, `r` and `d`, are described in Section 3.3. After finishing feature decomposition, the next call trains and calculates the performance of the generated model, using cross-validation (see Section 3.4):

```
./EDeN -i inputData.gspan.feature -f SPARSE_VECTOR -t inputData.target  
-a CROSS_VALIDATION -c 10
```

Cross-validation usually was set to 10-fold (`-c 10`). The target (`-t inputData.target`) file contains the class labels of the graphs (i.e. "1" for positive, "-1" for negative), which need to be in the same order as their corresponding graphs appear in the gSpan file. As a result, a prediction file is created, which can be evaluated by using `perf`²:

¹<http://www.bioinf.uni-freiburg.de/~costa/software.html>

²included in the EDeN archive

```
./perf -files inputData.target inputData.gspan.feature.predictions -ACC -SEN -ROC
```

The `perf` tool extracts performance measures such as AUROC, sensitivity or accuracy (see Section 3.5) from the prediction file. In this thesis, AUROC was used for comparing performances. In a different approach, the generated model is stored in a file and used later to measure its performance on a test dataset. First, the model gets trained and stored:

```
./EDeN -a TRAIN -i trainData.gspan.feature -f SPARSE_VECTOR -t inputData.target  
-m trainDataModel
```

Afterwards, model performance is tested by evaluating its capability to successfully classify test instances presented in a feature file:

```
./EDeN -a TEST -i testData.gspan.feature -f SPARSE_VECTOR -m trainingDataModel
```

The generated prediction file is then given to `perf`, together with the actual labels of the test instances, which evaluates the performance of the model on the test dataset.

A.3. Computational pipeline description

The computational pipeline roughly consists of three main scripts in the base directory, as well as six additional scripts in the `scripts/` subdirectory. Additionally, three shell scripts which control cluster job computation for the three main scripts can be found in the base directory:

```
01-generate-negative-sets.pl  
02-filter-sets-and-gspan.pl  
03-filter-gspan-and-run-eden.pl  
cluster-submit-01-generate-negative-sets.sh  
cluster-submit-02-filter-sets-and-gspan.sh  
cluster-submit-03-filter-gspan-and-run-eden.sh  
scripts/01-get-hsa-nm-list.pl  
scripts/02-seed-scan-nms.pl  
scripts/03-run-intarna-on-seed-hits.pl  
scripts/04-filter-intarna-hits.pl  
scripts/05-create-gspan-add-infos.pl  
scripts/06-filter-gspan-run-eden.pl
```

Table A.1.: Contents of the pipeline subdirectories.

Subdirectory	Content
<code>cluster-log/</code>	Log files for cluster computation
<code>data-tables/</code>	Dataset specific tables
<code>eden-files/</code>	EDeN calculation results
<code>gspan-results/</code>	gSpan files
<code>intarna-results/</code>	IntaRNA hybrid statistics files
<code>log-files/</code>	Log files generated by the pipeline scripts
<code>perl-lib/</code>	Perl pipeline library files
<code>programs/</code>	Binary files (EDeN, IntaRNA)
<code>results/</code>	Model performance results
<code>scanner-results/</code>	Negative seed scanning results
<code>scripts/</code>	Additional Perl scripts
<code>temp/</code>	Temporary files directory

All other files, including dataset tables and results files are stored in separate subdirectories. Table A.1 gives an overview of the subdirectory contents. Importantly, the scripts in the base directory utilize the scripts in `scripts/` which then accomplish the major computational tasks. Although they are not intended to be used other than in conjunction with the base directory scripts, the scripts can be executed directly from the base directory as well.

Each of the nine scripts contributes a help page which appears when the scripts are called with no parameters or the `-h` parameter. Importantly, in order to change the parameters and settings for gSpan filtering and EDeN, one has to change or add new entries to the gSpan filter and EDeN settings table. The table (`03-gspan-filter-list`) can be found in the subdirectory `data-tables/filter-tables/`. Each row of the table denotes the parameters which will be submitted to `scripts/06-filter-gspan-run-eden.pl` in one call. In case of enabled cluster computation and `n` table rows, the script will thus be called `n` times in parallel with `n` stated parameter lists.

Concerning the parameters, the script first needs to know which subgraph sections should be utilized for model training. This is accomplished by a set of parameters, from which exactly one has to be selected. The chosen parameter then determines which sections are included in the EDeN input graphs. Prior to filtering, the graph contains all sections and annotations as described in Section 4.3. Based on the selection, several section combinations are possible. For example, all sections or only the structure or the sequence sections can be retained. Beside the mandatory section selection, several optional filtering settings can be applied:

```
# Viewpoints settings:
-novp      Disable viewpoints
```

Appendix A. Computational Details

```
-vpseed      Use only seed region (1-8) as viewpoints vertices
-vpe x       Viewpoint extension, e.g. -vpe 10 for extending
              viewpoint + 10 on both ends

# Modify specific subgraph parts:
-minabs      Delete abstract sequence information in structure sections.
-stackex     Delete hybrid stacking information.
-strucex     Delete structure information in structure sections.
-hybrex      Delete hybrid information in structure sections.

# Filter positive instances by site quality or energy:
-ccppgg      <PERCENT> Cluster Cutoff Percentage Post Gspan Generation
              Based on normalized quality (coverage, enrichment).
-emaxpgg     <ENERGY> Give maximum IntaRNA energy allowed for IntaRNA hits
              Post gSpan generation filtering.

# For CLIP datasets:
-minclsh     Filter out CLIP clusters that overlap with CLASH clusters.
-minclsh25   Filter out CLIP clusters that overlap with CLASH clusters.
              Minimum overlap of 25 % required.
# For the combined CLIP dataset:
-minovlp     Filter out overlapping CLIP clusters inside the CLIP datasets.
-minovlp25   Filter out overlapping CLIP clusters inside the CLIP datasets.
              Minimum overlap of 25 % required.
```

In case of post gSpan generation site quality or energy filtering, seed types of the remaining positive instances are taken again to select the negative instances, as described in 4.4.2. Regarding the CLIP dataset filter options, filtering of clusters (target sites) that overlap with CLASH clusters was implemented in order to utilize CLASH as a test dataset (see Section 3.4). The following three `03-gspan-filter-list` row entries exemplify its usage:

```
1      -u -ccppgg 25 -r 3 -d 6
2      -s -ccppgg 25 -minabs -r 2 -d 6
3      -sp -ccppgg 25 -minabs -strucex -stackex -r 3 -d 5
```

Column one denotes the respective cluster array job ID, followed by gSpan filtering and EDeN parameters. The first three parameters (`-u`, `-s`, `-sp`) define the remaining subgraph sections for feature decomposition (`u`: sequence section only, `s`: the three shrep structure sections only, `sp`: three structure sections + protein information). All three jobs then apply site quality filtering (filter out worst 25 %), keeping only positive instances that reside upon the top 75 % sites. Regarding job 2 and 3, abstract structure information is also removed from the structure sections (`-minabs`). Additionally, job 3 removes structure and stacking information from the structure sections, leading to a conversion of the structure sections into three identical sequence sections with annotated protein crosslinks. The last two parameters (`-r`, `-d`) control EDeNs' feature decomposition process, as described in Section 3.3.

APPENDIX B

Abbreviations

(mi)RISC	(mi)RNA-induced silencing complex
miRNA	microRNA
3'UTR	3' untranslated region
5'UTR	5' untranslated region
AUROC	Area under the ROC curve
CDS	Coding sequence
CLASH	Crosslinking, ligation, and sequencing of hybrids
HITS-CLIP	High-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation
PAR-CLIP	3' Photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation
RBP	RNA-binding protein
SVM	Support Vector Machine

Declaration

Declaration of Academic Honesty

I hereby confirm that this thesis is my very own work and effort. Any other work described in this thesis has been properly cited and referenced.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Freiburg im Breisgau, February 13, 2014

Bibliography

- [1] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.
- [2] Julia Winter, Stephanie Jung, Sarina Keller, Richard I Gregory, and Sven Diederichs. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology*, 11(3):228–234, 2009.
- [3] Stefan L Ameres and Phillip D Zamore. Diversifying microRNA sequence and function. *Nature reviews molecular cell biology*, 14(8):475–488, 2013.
- [4] Matthias Selbach, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.
- [5] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105, 2009.
- [6] Danish Sayed and Maha Abdellatif. MicroRNAs in development and disease. *Physiological reviews*, 91(3):827–887, 2011.
- [7] Gunter Meister. Argonaute proteins: functional insights and emerging roles. *Nature Reviews Genetics*, 2013.
- [8] Dongmei Wang, Zhaojie Zhang, Evan O’Loughlin, Thomas Lee, Stephane Houel, Dónal O’Carroll, Alexander Tarakhovsky, Natalie G Ahn, and Rui Yi. Quantitative functions of Argonaute proteins in mammalian development. *Genes & development*, 26(7):693–704, 2012.

- [9] Markus Landthaler, Dimos Gaidatzis, Andrea Rothballer, Po Yu Chen, Steven Joseph Soll, Lana Dinic, Tolulope Ojo, Markus Hafner, Mihaela Zavolan, and Thomas Tuschl. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA*, 14(12):2580–2596, 2008.
- [10] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr, Anna-Carina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.
- [11] Alexander Maxwell Burroughs, Yoshinari Ando, Michiel Laurens de Hoon, Yasuhiro Tomaru, Harukazu Suzuki, Yoshihide Hayashizaki, and Carsten Olivier Daub. Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA biology*, 8(1):158–177, 2011.
- [12] Anne Dueck, Christian Ziegler, Alexander Eichner, Eugene Berezikov, and Gunter Meister. microRNAs associated with the different human Argonaute proteins. *Nucleic acids research*, 40(19):9850–9862, 2012.
- [13] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 2013.
- [14] Martin Reczko, Manolis Maragkakis, Panagiotis Alexiou, Ivo Grosse, and Artemis G Hatzigeorgiou. Functional microRNA targets in protein coding sequences. *Bioinformatics*, 28(6):771–776, 2012.
- [15] Ray M Marín, Miroslav Šulc, and Jiří Vaníček. Searching the coding region for microRNA targets. *RNA*, 19(4):467–474, 2013.
- [16] Shobha Vasudevan, Yingchun Tong, and Joan A Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–1934, 2007.
- [17] William Ritchie, Megha Rajasekhar, Stephane Flamant, and John EJ Rasko. Conserved expression patterns predict microRNA targets. *PLoS computational biology*, 5(9):e1000513, 2009.
- [18] Sooncheol Lee and Shobha Vasudevan. Post-transcriptional stimulation of gene expression by microRNAs. In *Ten Years of Progress in GW/P Body Research*, pages 97–126. Springer, 2013.
- [19] Wenqian Hu and Jeff Collier. What comes first: translational repression or mRNA degradation? the deepening mystery of microRNA function. *Cell research*, 22(9):1322–1324, 2012.

- [20] 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, 2012.
- [21] 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, 2012.
- [22] 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, 2010.
- [23] 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, 2011.
- [24] Thomas B Hansen, Trine I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, and Jørgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 2013.
- [25] Hun-Way Hwang, Erik A Wentzel, and Joshua T Mendell. A hexanucleotide element directs microRNA nuclear import. *Science*, 315(5808):97–100, 2007.
- [26] Kenji Nishi, Ai Nishi, Tatsuya Nagasawa, and Kumiko Ui-Tei. Human TNRC6A is an Argonaute-navigator protein for microRNA-mediated gene silencing in the nucleus. *RNA*, 19(1):17–35, 2013.
- [27] Andrew D Redfern, Shane M Colley, Dianne J Beveridge, Naoya Ikeda, Michael R Epis, Xia Li, Charles E Foulds, Lisa M Stuart, Andrew Barker, Victoria J Russell, et al. RNA-induced silencing complex (RISC) proteins PACT, TRBP, and dicer are SRA binding nuclear receptor coregulators. *Proceedings of the National Academy of Sciences*, 110(16):6536–6541, 2013.
- [28] Hongwei Liang, Junfeng Zhang, Ke Zen, Chen-Yu Zhang, and Xi Chen. Nuclear microRNAs and their unconventional role in regulating non-coding RNAs. *Protein & cell*, pages 1–6, 2013.
- [29] Donald E Kuhn, Mickey M Martin, David S Feldman, Alvin V Terry Jr, Gerard J Nuovo, and Terry S Elton. Experimental validation of miRNA targets. *Methods*, 44(1):47–54, 2008.
- [30] 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, 2009.

- [31] Daniel W Thomson, Cameron P Bracken, and Gregory J Goodall. Experimental strategies for microRNA target identification. *Nucleic acids research*, 39(16):6845–6853, 2011.
- [32] Lee P Lim, Nelson C Lau, Philip Garrett-Engele, Andrew Grimson, Janell M Schelter, John Castle, David P Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.
- [33] 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, 2009.
- [34] Alexander G Bartz, Mathias Munschauer, Björn Schwanhäusser, Alexandra Vasile, Yasuhiro Murakawa, Markus Schueler, Noah Youngs, Duncan Penfold-Brown, Kevin Drew, Miha Milek, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell*, 46(5):674–690, 2012.
- [35] Yoichiro Sugimoto, Julian König, Shobbir Hussain, Blaž Zupan, Tomaž Curk, Michaela Frye, and Jernej Ule. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol*, 13(8):R67, 2012.
- [36] Julian König, Kathi Zarnack, Gregor Rot, Tomaž Curk, Melis Kayikci, Blaž Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–915, 2010.
- [37] Lukasz Jaskiewicz, Biter Bilén, Jean Hausser, and Mihaela Zavolan. Argonaute CLIP – a method to identify in vivo targets of miRNAs. *Methods*, 2012.
- [38] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, 8(7):559–564, 2011.
- [39] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013.
- [40] 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, 2011.
- [41] Sung Wook Chi, Gregory J Hannon, and Robert B Darnell. An alternative mode of microRNA target recognition. *Nature structural & molecular biology*, 19(3):321–327, 2012.

- [42] Gabriel B Loeb, Aly A Khan, David Canner, Joseph B Hiatt, Jay Shendure, Robert B Darnell, Christina S Leslie, and Alexander Y Rudensky. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Molecular cell*, 2012.
- [43] 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, 2005.
- [44] 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 & development*, 18(2):132–137, 2004.
- [45] Dominic Didiano and Oliver Hobert. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature structural & molecular biology*, 13(9):849–851, 2006.
- [46] Yvonne Tay, Jinqiu Zhang, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124–1128, 2008.
- [47] 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. *Molecular cell*, 38(6):789–802, 2010.
- [48] 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. *Molecular cell*, 27(1):91–105, 2007.
- [49] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. Human microrna targets. *PLoS biology*, 2(11):e363, 2004.
- [50] Sabbi Lall, Dominic Grün, Azra Krek, Kevin Chen, Yi-Lu Wang, Colin N Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip MacMenamin, et al. A genome-wide map of conserved microrna targets in *c. elegans*. *Current biology*, 16(5):460–471, 2006.
- [51] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.
- [52] Ray M Marín and Jiří Vaníček. Efficient use of accessibility in microRNA target prediction. *Nucleic acids research*, 39(1):19–29, 2011.
- [53] Anke Busch, Andreas S Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 2008.

- [54] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10):1278–1284, 2007.
- [55] Dang Long, Rosalind Lee, Peter Williams, Chi Yu Chan, Victor Ambros, and Ye Ding. Potent effect of target structure on microRNA function. *Nature structural & molecular biology*, 14(4):287–294, 2007.
- [56] Stefan Ludwig Ameres, Javier Martinez, and Renée Schroeder. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, 130(1):101–112, 2007.
- [57] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, 8(1):69, 2007.
- [58] 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, 2005.
- [59] Ray M Marín and Jiří Vaníček. Optimal use of conservation and accessibility filters in microRNA target prediction. *PloS one*, 7(2):e32208, 2012.
- [60] Noa Bossel Ben-Moshe, Roi Avraham, Merav Kedmi, Amit Zeisel, Assif Yitzhaky, Yosef Yarden, and Eytan Domany. Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets. *Nucleic acids research*, 40(21):10614–10627, 2012.
- [61] Christelle Borel, Samuel Deutsch, Audrey Letourneau, Eugenia Migliavacca, Stephen B Montgomery, Antigone S Dimas, Charles E Vejnar, Homa Attar, Maryline Gagnebin, Corinne Gehrig, et al. Identification of cis-and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. *Genome research*, 21(1):68–73, 2011.
- [62] Tuuli Lappalainen, Michael Sammeth, Marc R Friedlaender, Peter AC’t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013.
- [63] Pål Sætrom, Bret SE Heale, Ola Snøve, Lars Aagaard, Jessica Alluin, and John J Rossi. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic acids research*, 35(7):2333–2342, 2007.
- [64] Andrea Rinck, Martin Preusse, Bernhard Laggerbauer, Heiko Lickert, Stefan Engelhardt, and Fabian J Theis. The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance. *RNA biology*, 10(6), 2013.

- [65] Ohad Balaga, Yitzhak Friedman, and Michal Linial. Toward a combinatorial nature of microRNA regulation in human cells. *Nucleic acids research*, 40(19):9404–9416, 2012.
- [66] Anja M Duursma, Martijn Kedde, Mariette Schrier, Carlos Le Sage, and Reuven Agami. miR-148 targets human DNMT3b protein coding region. *Rna*, 14(5):872–877, 2008.
- [67] 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 research*, 21(9):1395–1403, 2011.
- [68] 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 research*, 23(4):604–615, 2013.
- [69] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic acids research*, 32(16):4843–4851, 2004.
- [70] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNASHAPES: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.
- [71] Xifeng Yan and Jiawei Han. gSpan: graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 721–724. IEEE, 2002.
- [72] Steffen Heyne, Fabrizio Costa, Dominic Rose, and Rolf Backofen. GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, 28(12):i224–i232, 2012.
- [73] Daniel Maticzka, Sita Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *submitted*, 2013.
- [74] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262, 2010.
- [75] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [76] Ilho Ha, Bruce Wightman, and Gary Ruvkun. A bulged lin-4/lin-14 rna duplex is sufficient for caenorhabditis elegans lin-14 temporal gradient formation. *Genes & development*, 10(23):3041–3050, 1996.

- [77] 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, 2006.
- [78] Zhen Xia, Peter Clark, Tien Huynh, Phillippe Loher, Yue Zhao, Huang-Wen Chen, Isidore Rigoutsos, and Ruhong Zhou. Molecular dynamics simulations of Ago silencing complexes reveal a large repertoire of admissible 'seed-less' targets. *Scientific reports*, 2, 2012.
- [79] Guramrit Singh, Alper Kucukural, Can Cenik, John D Leszyk, Scott A Shaffer, Zhiping Weng, and Melissa J Moore. The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell*, 2012.
- [80] Liang Meng Wee, C Fabián Flores-Jasso, William E Salomon, and Phillip D Zamore. Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, 151(5):1055–1067, 2012.
- [81] Gavriel Mullokandov, Alessia Baccarini, Albert Ruzo, Anitha D Jayaprakash, Navpreet Tung, Benjamin Israelow, Matthew J Evans, Ravi Sachidanandam, and Brian D Brown. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nature methods*, 2012.
- [82] Jiayu Wen, Brian J Parker, Anders Jacobsen, and Anders Krogh. MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA*, 17(5):820–834, 2011.
- [83] Danny Incarnato, Francesco Neri, Daniela Diamanti, and Salvatore Oliviero. MREdictor: a two-step dynamic interaction model that accounts for mRNA accessibility and Pumilio binding accurately predicts microRNA targets. *Nucleic acids research*, 41(18):8421–8433, 2013.