Saarland University
Center for Bioinformatics
Bachelor's Program in Bioinformatics

Bachelor's Thesis

MiRNA Target Analysis Using Sequence Alignment

submitted by

Louisa Schwed

on September 21, 2016

Supervisor

Prof. Dr. Andreas Keller

Reviewers

Prof. Dr. Andreas Keller
Prof. Dr. rer. nat. Eckart Meese

Declaration

I hereby confirm that this thesis is my own work and that I have documented all sources used.

Saarbrücken, September 21, 2016

Louisa Schwed

Acknowledgement

I would like to express deep gratitude to my advisor, Prof. Dr. Andreas Keller, and his team for their effort and support for my bachelor thesis. Also, I would like to thank Prof. Dr. rer. nat. Eckart Meese for his time to review this thesis. Furthermore, I thank my friends and family for their encouragement and support during my studies and especially throughout the course of this thesis.

Abstract

Small endogenous RNAs, called miRNAs, are more and more subjects of investigation because they play an important role in the regulation of gene expression. After a process of maturation they are approximately 22 nucleotides long and target mRNAs by complementary binding. One of the many fields of research on miRNAs is the target prediction. The target prediction is a very complex problem and there is still no adequate solution for it. Many features like base pairing, free energy or conservation for a precise prediction can be incorporated. These features were found by investigation of known target interactions. The database miRTarBase provides a collection of experimentally validated miRNA target interactions.

In this research paper at hand, I will investigate these interactions with sequence alignments. MiRNA and respective target mRNA are aligned and the resulting binding sites are analysed further, especially with regard to how significant the alignment scores are compared to alignments of independent sequences. To analyse the significance of the scores, a negative control is generated by matching miRNAs with random gene sequences. With the help of a t-test and the two samples, p-values can be computed. With these values the significance can be estimated. The resulting low p-values show a significant increase in the alignment scores of true targets. However, considering the complementarity of the two aligned sequences, no big differences can be observed. The main sequence features, seed region towards the 5' end and additional 3' pairing, cannot be revealed. To get a more reliable prediction, the incorporation of additional features is recommended exemplified by existing prediction tools.

Contents

Chapter 1

Introduction

miRNAs are small, non-coding RNAs that function in the post-transcriptional regulation of gene expression, especially concerning gene silencing. A miRNA consists of approximately 22 nucleotides which build a single strand RNA [Bartel, 2004]. In a process of maturation the miRNA is processed for the actual binding mechanism. The whole maturation of the miRNA is shown in figure 1.1.

In a first step after transcription, the miRNA folds into a hairpin structure and builds the primary miRNA [Lee et al., 2002]. Then, this pri-miRNA is processed by Drosha which results in a precursor miRNA (pre-miRNA). This precursor is approximately 60 to 70 nucleotides long and contains two single stranded miRNAs. After the export of the miRNA into the cytoplasm another processing step is executed with Dicer to gain the actual form of a miRNA [Lee et al., 2002]. At this stage however, the miRNA is still double stranded and needs to be unwound to be a mature miRNA [Macfarlane and Murphy, 2010]. After processing of the initial double stranded miRNA-duplex, the guide strand will be further used in association with other molecules. In detail, the miRNA is mainly active in combination with a catalytic protein of the AGO protein family. With this protein, the miRNA builds an RNA-induced silencing complex [Ha and Kim, 2014]. This complex targets a mRNA mainly at its 3' untranslated region (UTR) by complementary binding to the sequence. In this case, the miRNA functions as a guide because its nucleotides pairs with the nucleotides of the target mRNA [Macfarlane and Murphy, 2010]. It can also bind at the 5' UTR or even in the coding region but these are rather uncommon regions. Either way, the binding results in gene silencing either by repression of mRNA translation or degradation of the respective mRNA [Enright et al., 2003]. Which one of the two pathways is chosen depends on the degree of complementarity. As figure 1.1 illustrates, if the complementarity is nearly perfect, the target mRNA is degraded. Otherwise, if the

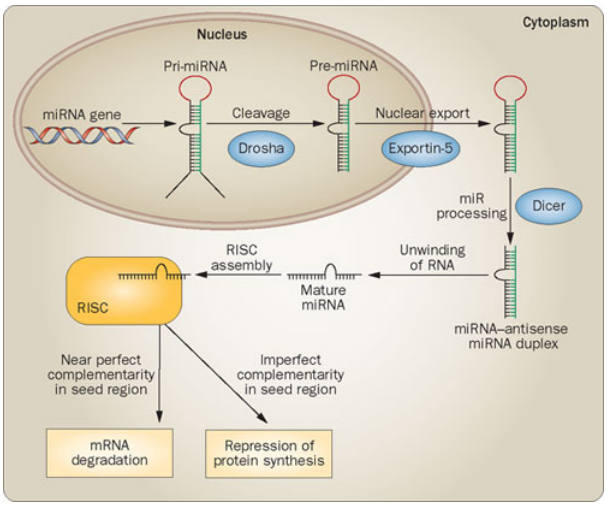

Figure 1.1: Biogenesis of miRNAs

seed region contains mismatches the translation is only repressed and not completely blocked. Both ways have a huge impact on the gene expression. The resulting genetic changes can have consequences like the initiation of cancer [Macfarlane and Murphy, 2010].

The gene regulation by miRNAs plays an important role for many major cell functions like growth, differentiation or metabolism and is currently examined by many researchers [Ardekani and Naeini, 2010]. The importance of miRNA and their regulation mechanism is also expressed by the fact that more than 60% of all human genes have at least one conserved miRNA-binding site [Friedman et al., 2009]. Therefore, a disregulation of miRNAs, either up- or downregulation, may result on the one hand in different human diseases, e.g. cancer or autoimmune diseases [Ardekani and Naeini, 2010]. On the other hand, this also leads to the assumption that miRNAs could be useful features for the diagnosis and treatment of diseases. However, they follow a very special and complicated targeting and this has to be elucidated. A first step is to understand their function and regulation mechanism. With this information it can be possible to predict targets of miRNAs and then specifically effect these target interactions to diagnose or treat diseases.

As already mentioned, especially this prediction of new targets is a major challenge in the whole field of miRNAs. There is still no adequate, reliable solution for the target prediction. Some tools exist that try to solve this problem. One of the first tools was miRanda which was developed by Enright et al. In their prediction algorithm they rely on three main features: sequence complementary, free energy and evolutionary conservation [Enright et al., 2003]. Another tool is miRSVR which uses bit vectors and

statistical learning approaches to predict targets. In the bit vector information like base complementarity, UTR length and distance, AU content and conservation is stored [Betel et al., 2010]. A third tool was developed by Bartel et al. and is called TargetScan. Their approach is based on seed matching and additional features like UTR positioning, AU content as well as base pairing of the nucleotides 12 or 13 to 16 of the miRNA. They especially take the seed region into account by defining four different patterns of seed binding sites [Lewis et al., 2005].

Because the miRNA binds to the sequence of its target gene an obvious starting point could be the complementarity of the miRNA sequence to the gene sequence which almost all of the tools consider at first. The first idea could be that if the sequence complementarity of miRNA and target mRNA is relatively high at a certain position, a new target site could be found. With the help of sequence alignments, these regions of high complementarity can be identified. In that case, very high alignment scores could indicate a high probability that there is a true binding site at this position. This alignment could then be used to predict completely new target sites and decide whether the considered mRNAs is targeted by a certain miRNAs. Whether this is a reliable indication or not will be discussed below. Before an assumption about using the alignment score as a feature can be made, there are some difficulties which have to be dealt with beforehand. Since it is known that the main binding happens within the seed region, other regions are not necessarily complementary and this may lower the complementarity and consequently the alignment score.

The main feature that most of the tools consider first is the sequence complementarity especially at a certain seed region. In contrast to miRNAs in plants which bind nearly perfectly complementary to their targets, miRNAs in animals bind less tightly and are not perfectly complementary [Rhoades et al., 2002]. There can be regions where the nucleotides are unbound which results in complex secondary structures which are hard to predict [Rehmsmeier et al., 2004]. The main complementary region or seed region includes nucleotides from position two to eight starting from the 5' end. Figure 1.2 shows the scheme of the seed region [Peterson et al., 2014].

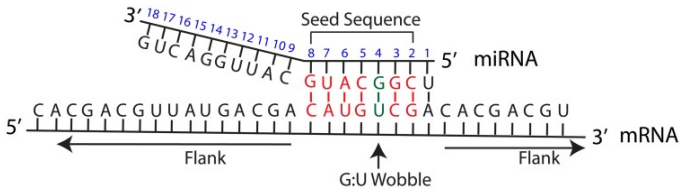

Figure 1.2: Scheme of seed matching

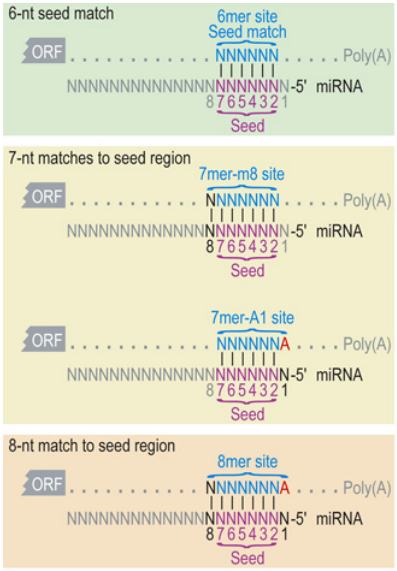

Figure 1.3: Canonical sites of seed region

The irregularity of the presence of non-complementary regions makes a reliable prediction for animals more difficult than for plants. Thus, concentrating mainly on the seed region for a prediction, will lead to many false positives. This small region of seven or eight nucleotides would be too unspecific because they can be present by chance in the mRNA although there is no real binding site at this position. It is recommended to take more than the seed region and the sequence complementarity into account to increase the complexity and specificity of the prediction. Additional features can be energies, conservation or site accessibility. How significant the consideration of the sequence complementarity of miRNA and mRNA sequence is, will be further analysed with validated data in this research. For these known interactions of the database the alignment scores are investigated in order to analyse whether these are significantly higher than for miRNA and sequences that are no real targets. Another point that is considered is my own found binding site accord with the ones given in the database that are provided by existing prediction tools.

In the mentioned seed region there can be different types of matching patterns: perfect matching between six nucleotides starting at the second nucleotide, seven nucleotides including the 8th position, seven nucleotides and an Adenine at position one or eight matching nucleotides from position one to eight and also an Adenine at position one [Grimson et al., 2007]. Figure 1.3 illustrates the different types of sites. Grimson et al. investigated the different types of sites referring to the effectiveness of the gene repression [Grimson et al., 2007]. Staying with sequence complementarity,

following a region of non complementary nucleotides towards the center of the miRNA, another matching region towards the 3' end can be observed. Aligning miRNA to target mRNA could shed light on the occurrence of this region.

To gain information about targeting and whether and to what extent alignment and sequence complementarity play a role in this process, databases provide reliable data about miRNAs and their targets. For these miRNA analyses, there are three main databases that provide the necessary information. A collection of experimental validated miRNA-target interactions(MTIs) is essential. The database miRTarBase provides a big dataset of validated MTIs [Hsu et al., 2011]. To get information about every single miRNA for different organisms miRBase contains large datasets about the miRNAs. The most interesting information is sequences of the precursor and of all mature miRNAs [Kozomara and Griffiths-Jones, 2011]. For further analysis I will only consider miRNAs of *Homo sapiens*. Last but not least, I used the UCSC Genome Bioinformatics Site. In their table browser you can access data files about human gene sequences and untranslated regions and their respective sequences [Karolchik et al., 2004].

Chapter 2

Data and Methodology

2.1 Data

2.1.1 miRTarBase

The database mirTarBase, which was released in 2010, provides by now about 7500 strong validated MTIs and 348000 weak ones from different species [Chou et al., 2016]. One constraint in this research is that I will focus on humans. Different experiment types were used to validate the data, including Reporter assay, Western plot, qPCR, Microarray, NGS, pSILAC and other methods whereas the first three are the ones that deliver strong evidence. The strong types detect the co-epression of miRNAs and their target gene, more precisely, the mRNA expression levels when miRNA is overexpressed or knocked down [Hsu et al., 2011]. Another constraint is that this thesis will focus on the strong evidence targets. The data collection provides a lot of information about the interaction between one miRNA and its target. Interesting details are the predicted alignments of miRNA and target 3' UTR sequence by either the author of the MTI or other prediction tools like miRanda. These alignments will also play an important role in the following analyses.

From the database catalogue the respective MTI data tables of the specific species can easily be downloaded, in this case *Homo sapiens*. The table contains the following information as shown in figure 2.1: miRTarBase ID, miRNA name, species of the miRNA, target gene symbol, target gene Entrez ID, species of the target gene, experiment type, support type, references. The fields of interest for the research are name of the miRNA, the target gene ID and the experiment type because it focuses on the Functional MTIs that are not weak. That means, every single strong validated interaction was analysed.

As mentioned above, the miRTarBase provides also binding sites as alignment positions predicted by different tools. To decide whether the alignments found in this research are compatible with the ones provided, the miRTarBase html page of the single MTI was parsed to get the start positions of each provided alignment. These positions do not exist for every miRNA but for about 3700 of the interactions. The parsed positions can then be compared to the resulting positions by the pairwise2 module. This will be described in detail under section 2.2 Methodology.

	B	C	D	E	F	G	H	I
1	miRNA	Species (miRNA)	Target Gene	Target Gene	Species (Target)	Experiments	Support Type	References
2	hsa-miR-20a-5p	Homo sapiens	HIF1A	3091	Homo sapiens	Luciferase reporter assay//Western blot//Northern blot	Functional MTI	18632605
3	hsa-miR-20a-5p	Homo sapiens	HIF1A	3091	Homo sapiens	Luciferase reporter assay//qRT-PCR//Western blot	Functional MTI	23911400
4	hsa-miR-20a-5p	Homo sapiens	HIF1A	3091	Homo sapiens	HITS-CLIP	Functional MTI (Weak)	22473208
5	hsa-miR-146a-5p	Homo sapiens	CXCR4	7852	Homo sapiens	qRT-PCR//Luciferase reporter assay//Western blot	Functional MTI	18568019
6	hsa-miR-146a-5p	Homo sapiens	CXCR4	7852	Homo sapiens	Microarray	Functional MTI (Weak)	20375304
7	hsa-miR-122-5p	Homo sapiens	CYP7A1	1581	Homo sapiens	qRT-PCR//Luciferase reporter assay	Functional MTI	20351063
8	hsa-miR-222-3p	Homo sapiens	STAT5A	6776	Homo sapiens	qRT-PCR//Luciferase reporter assay//Western blot	Functional MTI	20489169
9	hsa-miR-222-3p	Homo sapiens	STAT5A	6776	Homo sapiens	Luciferase reporter assay	Functional MTI	24736554
10	hsa-miR-21-5p	Homo sapiens	RASGRP1	10125	Homo sapiens	qRT-PCR//Luciferase reporter assay//Western blot	Functional MTI	20483747
11	hsa-miR-21-5p	Homo sapiens	RASGRP1	10125	Homo sapiens	Microarray	Functional MTI (Weak)	18591254
12	hsa-miR-21-5p	Homo sapiens	RASGRP1	10125	Homo sapiens	HITS-CLIP	Functional MTI (Weak)	22473208
13	hsa-miR-148a-3p	Homo sapiens	DNMT1	1786	Homo sapiens	Luciferase reporter assay	Functional MTI	20146264

Figure 2.1: Extract of the miRTarBase file

2.1.2 miRBase

To get the corresponding sequence of the miRNA name, miRBase, which was already published in 2005, provides a complete dataset of all known miRNAs [Griffiths-Jones et al., 2006]. By now it contains about 35000 mature forms and 2500 of them are found in *Homo sapiens*. The table with all miRNAs includes the accession number, miRNA ID, status, sequence, accession number of first mature form, its ID, its sequence, accession number of the second mature form, its ID and its sequence. I only use the miRNA ID and the respective sequences to align those to the gene sequence [Kozomara and Griffiths-Jones, 2011]. Figure 2.2 shows an extractof the file.

	A	B	C	D	E	F	G	H	I	J
1	Accession	ID	Status	Sequence	Mature1_Acc	Mature1_ID	Mature1_Seq	Mature2_Acc	Mature2_ID	Mature2_Seq
2	MI0000001	cel-let-7	UNCHANGED	UACACUGUGGAUCCGGUGAGG	MIMAT0000001	cel-let-7-5p	UGAGGUAGUAGGUUGUAUAGUU	MIMAT0015091	cel-let-7-3p	CUAUGCAAUUUUCUACCUUACC
3	MI0000002	cel-lin-4	UNCHANGED	AUGCUUCCGGCCUGUUCCUG	MIMAT0000002	cel-lin-4-5p	UCCCUGAGACCUCAAGUGUGA	MIMAT0015092	cel-lin-4-3p	ACACCUGGGCUCUCCGGGUACC
4	MI0000003	cel-mir-1	UNCHANGED	AAAGUGACCGUACCGAGCUGC	MIMAT0020301	cel-miR-1-5p	CAUACUUCCUUACAUGCCCAUA	MIMAT0000003	cel-miR-1-3p	UGGAAUGUAAAGAAGUAUGUA
5	MI0000004	cel-mir-2	UNCHANGED	UAAACAGUAUACAGAAAGCCAU	MIMAT0020302	cel-miR-2-5p	CAUCAAAGCGGUGGUUGAUGUG	MIMAT0000004	cel-miR-2-3p	UAUCACAGCCAGCUUUGAUGUGC
6	MI0000005	cel-mir-34	UNCHANGED	CGGACAAUGCUCGAGAGGCAG	MIMAT0000005	cel-miR-34-5p	AGGCAGUGUGGUUAGCUGGUUG	MIMAT0015093	cel-miR-34-3p	ACGGCUACCUUCACUGCCACCC
7	MI0000006	cel-mir-35	UNCHANGED	UCUCGGAUCAGAUCGAGCCAU	MIMAT0020303	cel-miR-35-5p	UGCUGGUUUCUUCCACAGUGGUA	MIMAT0000006	cel-miR-35-3p	UCACCGGGUGGAAACUAGCAGU
8	MI0000007	cel-mir-36	UNCHANGED	CACCGCUGUCGGGGAACCGCG	MIMAT0020304	cel-miR-36-5p	CGCCAAUUUUCGCUUCAGUGCUA	MIMAT0000007	cel-miR-36-3p	UCACCGGGUGAAAAUUCGCAUG
9	MI0000008	cel-mir-37	UNCHANGED	UUCUAGAAACCCUUGGACCAGI	MIMAT0015094	cel-miR-37-5p	UGUGGGUGUCCGUUGCGGUGCUA	MIMAT0000008	cel-miR-37-3p	UCACCGGGUGAACACUUGCAGU
10	MI0000009	cel-mir-38	UNCHANGED	GUGAGCCAGGUCCUGUUCCGG	MIMAT0020305	cel-miR-38-5p	UCCGGUUUUUUCCGUGGUGAUA	MIMAT0000009	cel-miR-38-3p	UCACCGGGAGAAAAACUGGAGU
11	MI0000010	cel-mir-39	UNCHANGED	UAUACCGAGAGCCCAGCUGAU	MIMAT0020306	cel-miR-39-5p	AGCUGAUUUCGUCUUGGUAAUA	MIMAT0000010	cel-miR-39-3p	UCACCGGGUGUAAAUCAGCUUG
12	MI0000011	cel-mir-40	UNCHANGED	UCCUGUCCGCACCUCAGUGGA	MIMAT0020307	cel-miR-40-5p	AGUGGAUGUAUGCCAUGAUGAUA	MIMAT0000011	cel-miR-40-3p	UCACCGGGUGUACAUCAGCUAA
13	MI0000012	cel-mir-41	UNCHANGED	GGGUCCCAGAGACCUUGGUGG	MIMAT0020770	cel-miR-41-5p	GGUGGUUUUUCUCUGCAGUGAUA	MIMAT0000012	cel-miR-41-3p	UCACCGGGUGAAAAAUCACCUA

Figure 2.2: Extract of the miRBase file

2.1.3 UCSC Genome Bioinformatics Site

The last required dataset is the collection of target gene sequences and their respective untranslated regions (UTRs). On the UCSC Genome Bioinfor-

matics Site I generated a list of all genes and their UTRs of the human genome using the Table Browser [Karolchik et al., 2004]. The selected assembly was Dec. 2013 (GRCh38/hg38) and the track was RefSeq Genes. As an output format I selected the sequence. Next, the sequence type was set to genomic and the regions 5' UTR Exons, 3' UTR Exons and introns were chosen. The resulting list consists of a description of the gene with the transcript accession number (NM-number) and the concatenated sequences of 5' UTR, gene and 3' UTR as shown in figure 2.3.

```
>hg38_refGene_NM_001276352 range=chr1:67092176-67134971 5'pad=0 3'pad=0 strand=- repeatMasking=none
cgagtaacccgcggagccagaagagggaggaaaggagatgagatttcatc
atgttggccagcctggtctcaaactcctgacctcaagtgacccgcctgcc
tcagcctcccaaagtgctgggattacaggaatttagtgattgacaATGGC
AGAAAAAATCCTAGAGAAGTTGGATGTCCTTGATAAGCAAGCAGAGATAA
TCTTGGCCAGAAGAACAAAGATAAACAGGCTTCAGAGTGAAGGAAGAAAA
ACAACTATGGCTATACCCCTGACATTTGATTTTCAGTTGGAATTTGAAGA
...
```

Figure 2.3: Extract of the UCSC genome file

2.1.4 ID converter

For the alignment of miRNA to the corresponding gene the respective miRNA sequence and gene sequence are required. The interaction data from miRTarBase only provides the correlation between miRNA name and target gene ID. In contrast, the dataset from UCSC only delivers the NM number for the gene sequence. Therefore a conversion from gene Entrez ID to Refseq mRNA accession number is required. Biodbnet provides a conversion tool for different IDs, names and numbers [Mudunuri et al., 2009]. I entered all existing target Entrez IDs of the miRTarBase file and obtained a conversion list of IDs and their corresponding Refseq accession numbers. In my program I used this list by storing every entry in a dictionary and looked up the particular ID for every MTI. This is also described in the following section.

2.2 Methodology

The program for the analysis is implemented in Python. As input files the UCSC gene file, miRTarBase file and miRBase file are required in a csv format. Additionally the conversion file of gene ID and NM number is needed. It can be generated as described in section 2.1.4 ID converter and used as an input file in a text format. For each data file, except the MTI file, an own dictionary is generated. Therefore the data files are parsed and each entry is stored into the respective dictionary that allows fast searching. For

the miRNAs the dictionary is defined as follows: miRNA name as key and sequence as value. The two mature forms, in case there exist two, are stored separately. For the gene file, the NM number is the key and the value is a list of sequences of 5' UTR, coding region and 3' UTR. At last, for the conversion list, the Entrez ID is stored as the key and the list of NM numbers as a value. More than one NM number for one gene Entrez ID can exist because different transcripts can be obtained by alternative splicing.

Then for every MTI in the miRTarBase file the respective miRNA sequence is searched for in the miRBase dictionary and combined with the gene sequence of the corresponding target in the gene dictionary. The miRTarBase file contains data of 'Functional MTI', 'Functional MTI(Weak)' and 'Non-Functional MTI'. In this paper, I only consider the strong validated data of the Functional MTIs because they deliver strong evidence for the interaction. Hence, if an entry in the miRTarBase file contains something other than 'Functional MTI', it is skipped.

After matching miRNA and gene, a local alignment is performed. Therefore, I took the whole gene sequence, meaning the concatenation of UTR and coding region. This will show whether the preferred binding will be within the 3' UTR, as known, or somewhere else. An alignment is defined as the optimal positioning of the bases of one sequence, in this case the miRNA, to a region in the other sequence, the gene sequence. The result reveals functional or structural similarities. [Bioinformatics.org, 2010]. In this case, the similarity can be interpreted as a region where a binding site could be present. Therefore, if we take the reverse complement of the miRNA sequence and align it to the mRNA sequence, this will simulate the binding. More precisely, if the alignment score is high, the sequences (mRNA and reverse complement of miRNA) are more similar, implying the actual miRNA could possibly bind at this alignment position. Because the miRNA can bind at any region in the gene, a local, not a global, alignment is executed.

The Biopython library provides a module, pairwise2, for pairwise local alignments of two sequences. This tool is based on a dynamic programming algorithm. This function can be used either with default parameter or different scores and costs can be defined. The default parameters are as following: +1 for matching character, 0 for not matching ones and there are no gap penalties [Biopython, 2016]. To get a more suitable alignment, own parameters can be selected. Table 2.1 shows the different parameters I used to analyse the data. Set no. 8 is similar to the parameters they used for the tool miRanda [Enright et al., 2003]. The other parameters are logically selected to test which influence they have on the results.

Table 2.1: Parameter sets

Parameterset	match score	mismatch score	gap open	gap extend
1	default: 1	default: 0	-4	-4
2	default: 1	default: 0	-5	-1
3	1	-2	-2	-1
4	2	-2	-5	-4
5	3	-2	-4	-4
6	5	-1	-8	-4
7	5	-2	-8	-3
8	5	-3	-8	-2
9	5	-4	-6	-4

For each alignment the following steps are performed: the pairwise2 module delivers the best alignment with the highest score. Sometimes there is more than one alignment because they have the same highest score. For each found alignment list, the miRNA name, the respective NM-number and the alignment score are stored in a string. Then, for each single alignment in the list, the starting position is computed and added to this string. To get a final list of all alignments, the single strings of each MTI is written to a text file.

For the analysis of the complementarities, the alignment is further examined. First, both sequences are reversed because they are given from 3' to 5'. To further analyse the 5' seed region and make the plots more clear, the sequences then start with the 5' end. Each nucleotide position of miRNA and target mRNA is compared and whenever the nucleotide is the same, a '1' is added to a list, otherwise, if they are different or there is a gap, a '0' is added. This list is yielded to a big matrix for the final analysis. Going through this matrix through each column the number of '1's is computed and divided by the total number of elements in the matrix. The result will be the ratio of complementary bases. This ratio is plotted for each position in the alignment.

To be able to analyse the data statistically a set of negative controls is required. To produce this data, 1000 miRNAs of the miRBase file were randomly assigned to a list of genes. Then, this non target file is processed in the same way as the file of the true target interactions. The random assignment can by chance contain true targets but generally the miRNAs are matched with genes that are not targeted by them.

For the datasets the average and standard deviation were computed with functions of LibreOffice. A statistical two-tailed t-test for two samples with

equal variances was performed with LibreOffice as well [King, 2013]. In this case, one sample was the data set of the true targets, the other one the data of the non-targets. The t-test function provides a p-value as a result. The lower this p-value is, the more significant is the increase of the alignment score of the true targets. If there was an increase in the score, it would indicate a true target. Hence, if we only calculate the alignment score for a new miRNA and any target sequence and observe a high score, the probability that a new, true target can be found would be high. Whether this is a reliable feature for the prediction will be discussed in the following.

For the analysis of the alignment starting position, the miRTarBase was parsed first. For each strong validated data in the file the respective miRBase target page was parsed and the three given starting positions were searched. In order to do that, the MTI ID was inserted into the url which was opened. Together with the corresponding miRNA name and target NM number the positions were stored in a text file. In the final table the starting positions found were compared to the positions given in the resulting text file what was mentioned before. The final number of positions found is stored in the final result table as well.

Chapter 3

Results

For each parameter set the computed alignment score for every MTI is stored in a table. Additionally the average alignment score, the standard deviation and the p-values are listed below. Table 3.1 shows only a summary of the resulting table of the program, excluding the single alignment scores of each MTI that the program provides. 13934 alignments were computed for the true targets and 2717 for the non-targets. To draw a better comparison, the scores of the non-targets are listed right under the scores of the true targets. The averages of both alignment score sets show that the scores of the true targets are in general only slightly higher than the ones of the non-targets. For the lower alignment scores e.g. of set 1 -2 -2 -1, resulting from really low match scores, the difference is only 0.3, and at first sight not very significant. For the higher scores, the difference amounts to two which is an increase of about 3% for the parameter set 5 -1 -8 -4. At first sight, it is hard to say if this is significant because the scores are generally higher and therefore the amount of two is not much.

The standard deviation of the non-targets is slightly higher than that for the true targets but also not very significant. Hence, the scores of both cases are similarly spread from their means. The p-value sheds light on whether the increase in the alignment score is significant when considering the population of scores. According to the low resulting p-values of the t-test, which were computed as described in section 2.2 Methodology, the increase in the alignment scores is significant. The parameters 5 -1 -8 -4 show the most significant difference with a p-value of $2.529 * 10^{-61}$ whereas 2 -2 -5 -4 shows the least significant increase but also here the p-value is very low with $3.279 * 10^{-36}$. So even though the average scores do not differ by much, the scores of the true targets seem to stem from the higher scores and the ones of the non-targets from the lower scores. Additionally, there are many more scores for the true targets than for the negative control. If the average score is about two points higher for this parameter set, the bigger data set of true

Table 3.1: Table of alignment results

	3 -2 -4 -4	5 -1 -8 -4	5 -2 -8 -3
Average true targets	33.096	64.247	59.495
Average non-targets	31.831	62.199	57.464
Standard deviation true targets	4.058	5.758	6.304
Standard deviation non-targets	4.233	6.510	6.753
t-test p-value	$5.934\ 10^{-49}$	$2.529\ 10^{-61}$	$1.040\ 10^{-51}$

	5 -3 -8 -2	5 -4 -6 -4	-4 -4
Average true targets	56.391	55.510	13.712
Average non-targets	54.269	53.294	13.362
Standard deviation true targets	6.633	6.806	1.078
Standard deviation non-targets	6.876	7.01	1.311
t-test p-value	$1.394\ 10^{-51}$	$1.744\ 10^{-53}$	$8.824\ 10^{-50}$

	-5 -1	1 -2 -2 -1	2 -2 -5 -4
Average true targets	13.710	8.593	18.351
Average non-targets	13.361	8.208	17.619
Standard deviation true targets	1.078	1.347	2.751
Standard deviation non-targets	1.309	1.385	2.868
t-test p-value	1.13110^{-49}	$8.157\ 10^{-42}$	$3.279\ 10^{-36}$

targets has to include many more MTIs with higher alignment score to increase it by two. Taking this consideration into account, the amount of two is not that low. Hence, there must be many alignments with higher scores.

Another observation is that the difference between the parameters is not that big. For each set, the increase is still very significant, even though the p-values range from 10^{-36} to 10^{-61}. There is no exception that does not show significance in the alignment scores of the true targets. The differences in the value of the scores in general are obviously a result of the different match and mismatch scoring and penalizing gaps which can be more or less. A certain penalty for gaps should be present because it is not very logical from a biological sight of view that the alignment contains many gaps. This alignment as described above should simulate the binding to the target binding site. If a miRNA binds to its target, it binds with some regions of complementarity as well as with some bulges. However, a binding with too many bulges is not common and the regulation would not function very

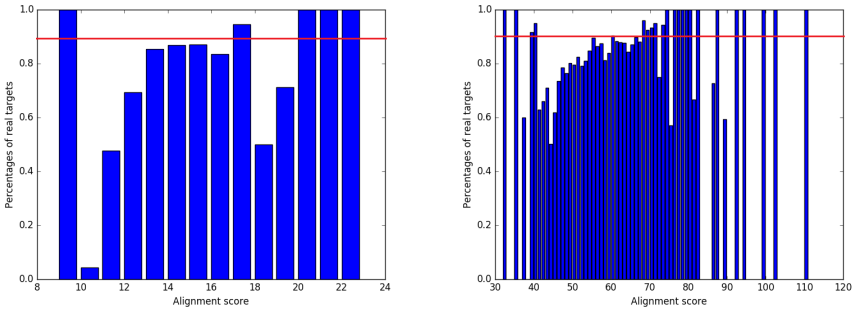

Figure 3.1: Distribution of alignment scores, parameters -4 -4 and 5 -3 -8- 2

well. Therefore few gaps are reasonable but many and longer ones should not be present. The parameter set with the highest p-value has low match and mismatch scores compared to the gap costs. So gaps are more penalized than mismatches. This will result in binding sites with less gaps and some mismatches. The complementarity of the true targets will therefore be lower and the binding site will not be very specific. Analysing the non-targets, they should not have true binding sites because they are randomly assigned to miRNAs. Therefore, aligning these two sequences should generally result in worse binding sites where the complementarity is as well not very high. Comparing them to the true targets which have also worse binding sites because of the scoring, the difference between the two groups will not be very high. And this can be seen in the higher p-value. Nevertheless it is still significant.

The increase in the alignment score can also be observed when considering the distribution of the scores within the two groups of case and control. For two parameter sets (No. 1 and No. 8) the distribution was plotted and is shown in figure 3.1. On the x-axis the different existing scores of the respective data set are shown and on the y-axis the ratio of true targets that had these scores. It can be observed that towards the higher alignment scores the ratio of true targets rises. However, there are some scores that are exceptions because more non-targets produced these scores than true targets, shown by lower y-values. Another reason for low y-values or values of 0 can be that neither true targets nor non-targets produced them. Other deviations can be due to not considering the differences in the number of alignments between true targets and non-targets. To achieve more precise results this should be incorporated.

Using this score as a prediction feature, a threshold needs to be determined to identify a true target. Hence, a certain ratio has to be exceeded to rank this score as high enough for being a true target. In figure 3.1 a threshold of 90% is delineated meaning at least 90% true targets produced

a certain alignment. Still 10% of random non-targets had this score as well. Assuming this threshold and using the alignment scores for a prediction the first parameter set shows that if we have scores of nine, 17, 20, 21, 22 or 23, a mRNA can be classified as a true target. On the one hand, obviously the score nine would not fit in the pattern because it is a very low alignment score and the binding site would not be very suitable because there certainly are many mismatches or gaps that lower the alignment score. On the other hand, alignments of miRNA and mRNA with a resulting score of 19 which belongs to the higher scores for this set, would not be classified as a true target even though the binding might be suitable. How this classification will look like strongly depends on this threshold. A lower threshold will result in a lower specificity and higher sensitivity and vice versa for the higher one. The two measurements of the performance need to be balanced and determined depending on the analysis.

For the second plot in figure 3.1 the same can be observed: some lower alignment scores are above the threshold although they are not suitable and for some higher ones the probability to be a true target would be low. But in general it can be said that if the score is above 70, the likelihood that the target is a true target is quite high. In this plot it can also be seen that there are some outliers with high alignment scores. There are very few with rather high scores like 100 and 110 but scores in between the two were not produced, neither by true targets nor by non-targets.

Summing up, assuming the score as the only prediction feature and given a certain threshold, some mRNAs are predicted right to be true targets but there are also false positives and false negatives. Given high scores of the alignment of miRNA seqeunce and mRNA sequence, a mRNA is not necessarily targeted by this miRNA. This uncertainty can be accounted for by considering and combining many more features such as conservation or energies and get a more reliable prediction.

As described in the Methodology, for each parameter set the complementarity of the single alignment positions of the miRNA and target gene sequence were plotted. These plots for the different sets are shown in figure 3.2. The x-axis of the plots shows the position one to 22 of the miRNA, the y-axis represents the ratio of bases that are complementary at this position, considering all alignments produced with these parameters. Because most of the miRNAs are only 22 nucleotides long, I will only focus on these positions. Higher positions would be very imprecise because of the different lengths. The non-targets are shown on the right site, the true targets on the left.

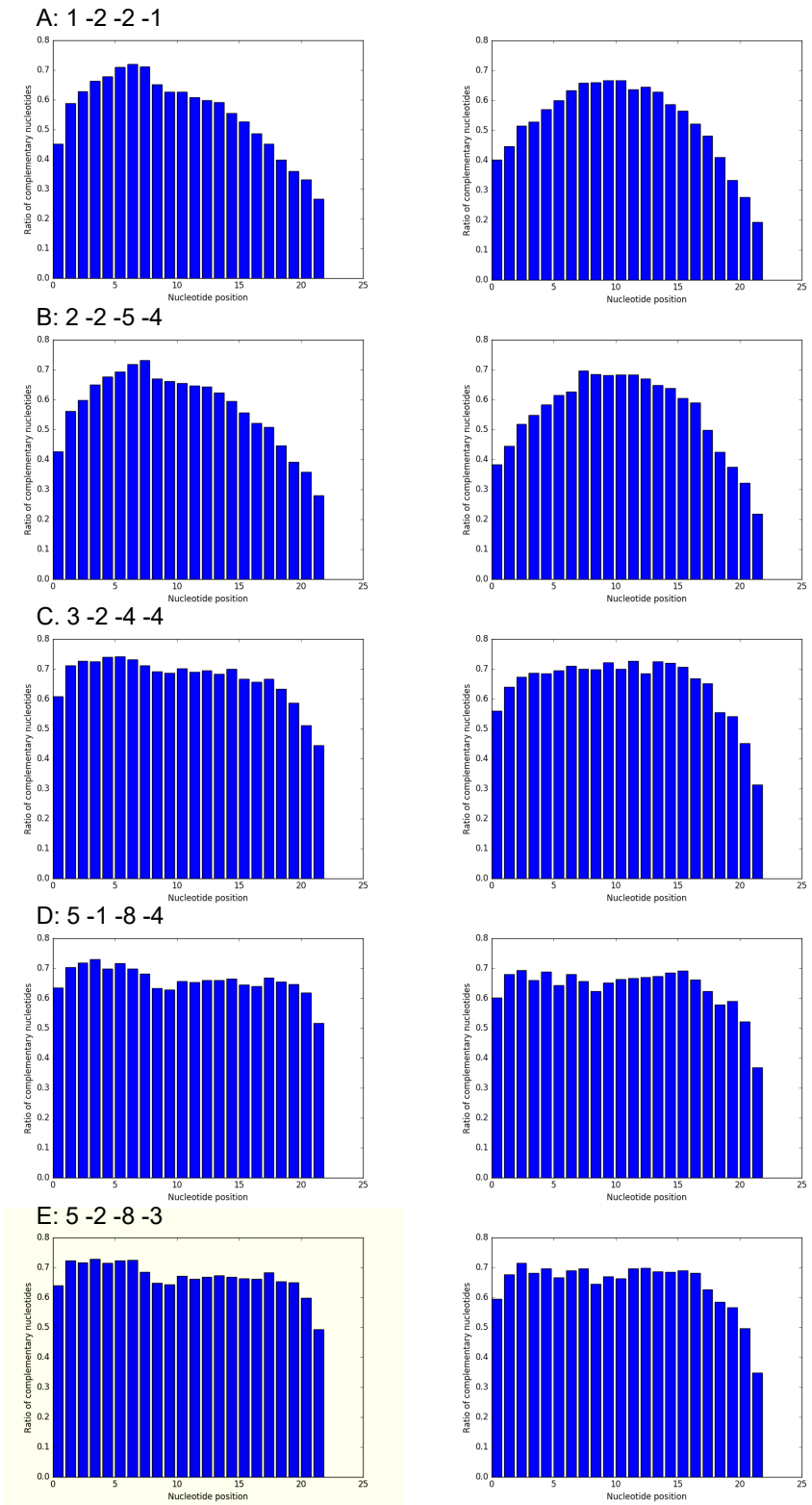

Figure 3.2: Ratios of complementarity per parameter set (A-I): targets left, non-targets right

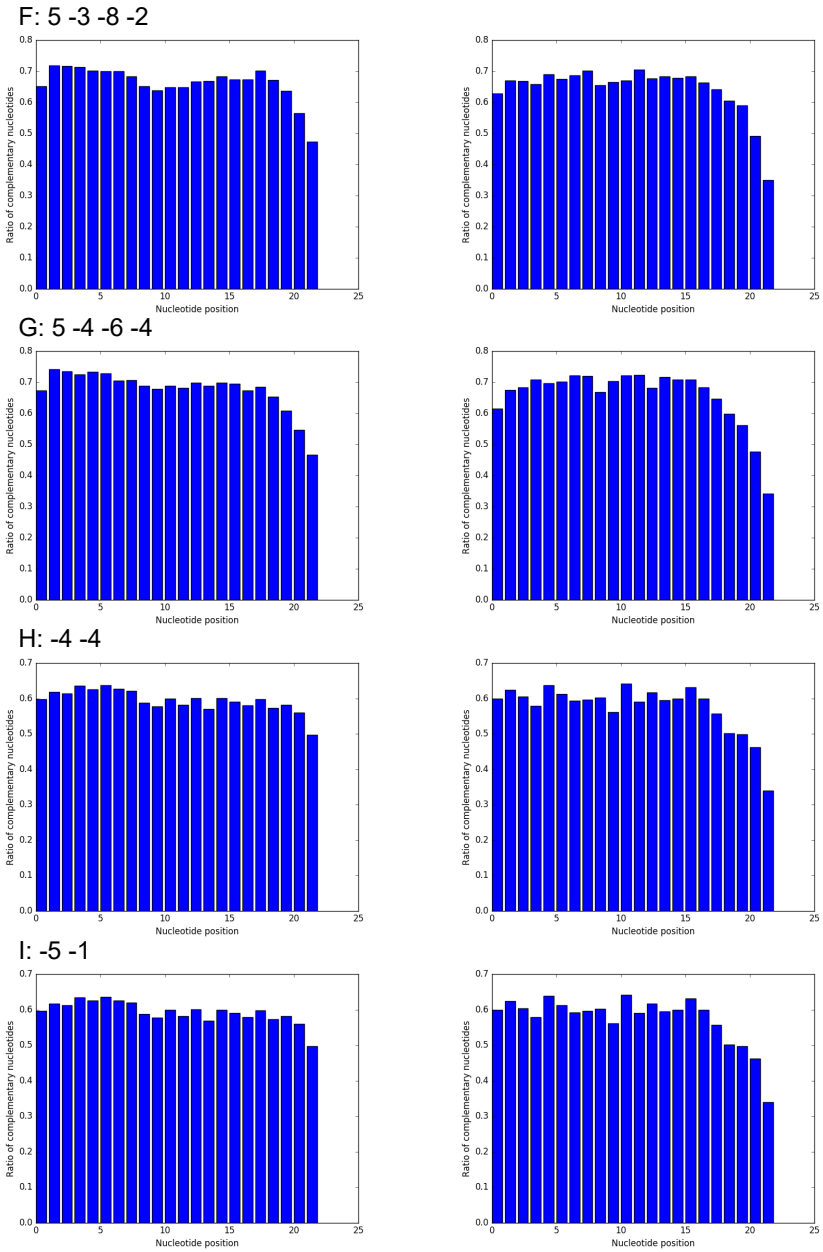

Figure 3.2 continued

In general it can be observed that the bars of the non-targets are never higher than 70%. The true targets sometimes exceed this percentage but not very often. For some sets there is a more significant difference between true targets and non-targets than for others. Regarding set A, for the true target left an increase in complementarity in the seed region can be observed. Whereas the first position is below 50%, the complementarity rises from the second position on to the 8th. The ratio towards the end gets really low down to 30% of complementary bases. In comparison, the maximum of complementarity of the non-targets right is shifted towards the central positions, not showing the typical seed region. From then on it decreases very fast as well. Notably the plot of the true targets does not show the complementary region towards the 3' end. The same applies for plot B. Although the parameters are very different the plot for the true targets looks similar to that of A. The right plot even more reveals the difference between the two groups. There is again an increase in the positions eight to 13, not fitting into any rule of targeting.

Considering the other figures of the complementarities, the seed region is always recognizable, although not very strong. The increase in the first few nucleotides amount only to a few percentages. The plots of the non-targets do not follow any rules as expected. The complementarities are almost constant and drop at the end down to 30%. For the true targets the decrease at the end can also be observed but not so much. Plot H and I stay almost constant above 50%.

Summing up for the true targets the complementarity is almost as high as for the non-targets, gaining no reliable information for the prediction. This implies that the consideration of the complementarities of the miRNA positions is not a significant and reliable prediction feature. Only having these ratios, it can not certainly be decided whether this plot shows a true targeting or not.

In figure 3.3 three plots with curves of three parameter sets illustrate the complementarities of the nucleotides in an alternative way. In the first plot the difference in the seed region can be seen. The blue curve of the true targets is about 10% higher than the curve of the non-targets. From nucleotide eight on, the red curve declines from its maximum until it crosses the blue one but the difference is not very big. From then on the curves are very similar and they both drop down to 30%. So for this parameter set only a difference in the seed region can be observed. The second plot of the parameter set 5 -4 -6 -4 shows even less differences. Again, there is a small difference in the seed region where the blue curve is slightly higher but only up to 5%. Then both curves stay at a high level between 65 and 70% which is a contrast to the first plot. But the main point, a difference between true

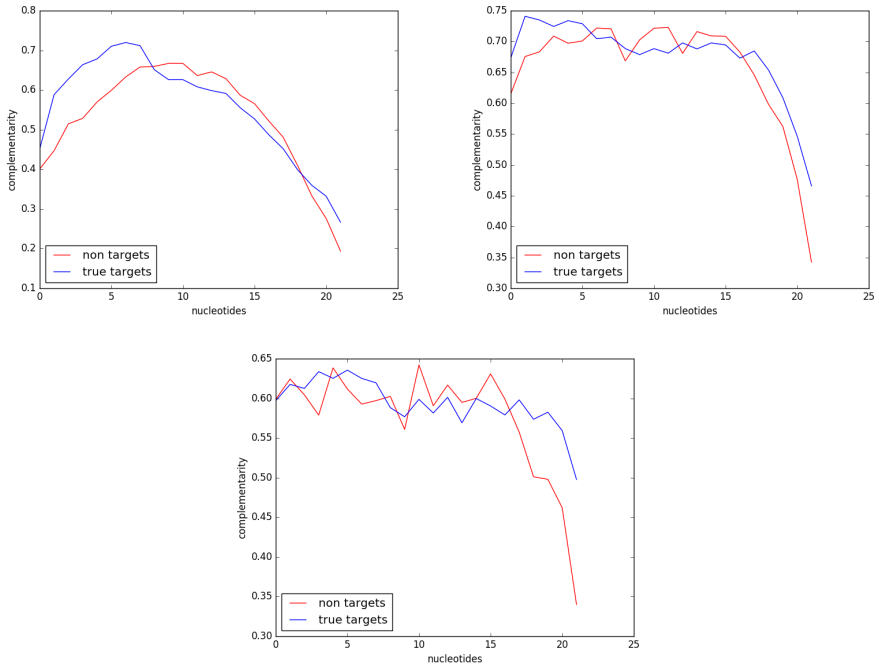

Figure 3.3: Curves of complementarity of non and true targets: 1 -2 -2 -1, 5 -4 -6 -4 and -5 -1

and non-targets can not be observed. The last plot is slightly different from the two others. The blue and the red curve alternate and are very similar in height. In this graph, neither seed region nor additional 3' pairing can be observed. Similarly, additional pairing cannot be observed at all for the two other graphs. All in all, the plots and the complementarity of the pairing nucleotides deliver not much information. Concluding, having a look at the complementarity of a new examined binding site is not significant to decide whether this position is a newly found binding site and hence an actual target of a miRNA.

The table in figure 3.4 shows that the average of the complementarity ratios are similar when comparing true targets with non-targets. The nucleotides of the seed region are highlighted in blue and the ones of the additional 3' pairing in orange. Calculating the average complementarity ratios of these positions sheds light on the presence of these regions in validated interactions. For the first parameter set, the true targets show a small increase in the seed region but a decrease in the additional pairing. This feature of additional pairing can not be significantly observed. For the same set and its negative control, a constant average can be seen. Almost all true targets show an increase in the seed region and a decrease in the

Nucleotides	-4 -4	-5 -1	1 -2 -2 -1	2 -2 -5 -4	3 -2 -4 -4	5 -1 -8 -4	5 -2 -8 -3	5 -3 -8 -2	5 -4 -6 -4
True targets									
1	0.598	0.597	0.452	0.427	0.608	0.634	0.641	0.652	0.673
2	0.619	0.618	0.588	0.562	0.712	0.703	0.724	0.718	0.741
3	0.614	0.613	0.628	0.599	0.727	0.718	0.717	0.716	0.735
4	0.635	0.634	0.664	0.650	0.725	0.730	0.729	0.713	0.724
5	0.626	0.625	0.678	0.676	0.740	0.698	0.716	0.701	0.734
6	0.637	0.636	0.710	0.694	0.742	0.716	0.723	0.701	0.729
7	0.627	0.625	0.720	0.718	0.732	0.699	0.724	0.700	0.705
8	0.621	0.620	0.712	0.731	0.712	0.682	0.685	0.683	0.707
9	0.588	0.588	0.651	0.669	0.692	0.633	0.649	0.652	0.689
10	0.577	0.577	0.626	0.662	0.687	0.628	0.643	0.638	0.679
11	0.599	0.599	0.626	0.655	0.701	0.657	0.672	0.649	0.689
12	0.582	0.581	0.608	0.647	0.690	0.654	0.662	0.648	0.681
13	0.601	0.601	0.598	0.643	0.696	0.659	0.668	0.666	0.698
14	0.570	0.569	0.591	0.624	0.683	0.660	0.673	0.669	0.688
15	0.600	0.600	0.555	0.595	0.700	0.666	0.669	0.683	0.698
16	0.590	0.590	0.527	0.557	0.667	0.645	0.664	0.673	0.695
17	0.580	0.579	0.487	0.522	0.657	0.641	0.661	0.673	0.673
18	0.598	0.598	0.452	0.508	0.666	0.668	0.684	0.702	0.685
19	0.573	0.574	0.399	0.447	0.634	0.655	0.654	0.672	0.654
20	0.582	0.583	0.360	0.392	0.586	0.647	0.649	0.636	0.609
21	0.560	0.559	0.332	0.358	0.512	0.619	0.598	0.565	0.547
22	0.497	0.498	0.267	0.280	0.444	0.516	0.494	0.473	0.466
Average	0.594	0.594	0.556	0.573	0.669	0.660	0.668	0.663	0.677
Average 2 – 8	0.625	0.624	0.671	0.661	0.727	0.707	0.717	0.704	0.725
Average 13 – 16	0.591	0.590	0.568	0.605	0.686	0.658	0.669	0.673	0.695

Negative control									
1	0.599	0.599	0.401	0.383	0.560	0.601	0.595	0.628	0.615
2	0.625	0.624	0.447	0.445	0.639	0.679	0.676	0.671	0.676
3	0.605	0.604	0.515	0.518	0.674	0.693	0.715	0.668	0.683
4	0.579	0.579	0.529	0.549	0.686	0.660	0.681	0.658	0.709
5	0.638	0.638	0.570	0.583	0.685	0.688	0.696	0.689	0.697
6	0.612	0.612	0.599	0.614	0.695	0.643	0.666	0.675	0.701
7	0.593	0.593	0.633	0.627	0.710	0.680	0.690	0.687	0.722
8	0.597	0.597	0.658	0.697	0.701	0.657	0.697	0.701	0.721
9	0.603	0.602	0.660	0.685	0.698	0.624	0.646	0.654	0.669
10	0.561	0.561	0.667	0.682	0.722	0.651	0.671	0.665	0.703
11	0.642	0.642	0.667	0.683	0.700	0.663	0.663	0.671	0.722
12	0.590	0.591	0.637	0.683	0.727	0.667	0.697	0.705	0.723
13	0.617	0.617	0.646	0.671	0.685	0.670	0.698	0.677	0.681
14	0.595	0.595	0.629	0.649	0.726	0.674	0.687	0.684	0.716
15	0.600	0.600	0.587	0.639	0.720	0.684	0.684	0.678	0.709
16	0.631	0.631	0.565	0.606	0.707	0.692	0.691	0.684	0.709
17	0.599	0.599	0.522	0.590	0.668	0.662	0.681	0.663	0.683
18	0.558	0.557	0.481	0.498	0.652	0.623	0.626	0.642	0.646
19	0.502	0.501	0.410	0.426	0.555	0.578	0.584	0.605	0.598
20	0.498	0.498	0.333	0.376	0.542	0.589	0.567	0.590	0.562
21	0.462	0.462	0.276	0.321	0.451	0.522	0.496	0.491	0.477
22	0.340	0.340	0.194	0.219	0.314	0.369	0.349	0.350	0.342
Average	0.575	0.575	0.528	0.552	0.646	0.635	0.643	0.643	0.657
Average 2 – 8	0.607	0.607	0.564	0.576	0.684	0.671	0.689	0.679	0.701
Average 13 – 16	0.611	0.611	0.607	0.641	0.710	0.680	0.690	0.681	0.704

Figure 3.4: Table of ratios of complementarities

other region. For the non-targets there are different patterns. For example, for the third parameter set both regions increase compared to the standard average. For the other sets there is either no big difference or both increase but never decrease. Referring to the plots, the seed region in the true targets is slightly more visible than for the non-targets. All in all, the differences are not very big and significant. Both regions cannot be observed as clear as expected.

The analysis of the alignment starting positions was not particularly successful and significant. Table 3.2 shows the numbers of the own predicted positions that were also given in the miRTarBase. It can be seen that two sets deliver only half the number of the other ones (5 -2 -8 -3 and 5 -3 -8 -2). These two sets also use very similar parameters. That shows that the alignment positions strongly depend on the selected parameters. In comparison to the given number of about 3700 positions in the miRTarBase, the numbers of the respective own found ones are in general not very high. Only about 20% were found. This could be due to the different sizes of the UTRs of the genes. In this research the UTRs are from the UCSC website whereas the miRTarBase might use another source. Therefore the sizes can be a bit different resulting in different defined positions, although the alignment would start at the same position. To eliminate these small deviations, I allowed a window of 10 positions where the starting position can be. So if the provided miRTarBase position is given, the own predicted position is classified as consistent with the provided one if the position lies in a window of +5 or -5 related to the miRTarBase position. Another point is the difference that I align the miRNA sequence to the whole gene sequence. The best alignment position can therefore lie within the other region whereas the positions of the miRTarBase are only in the 3' UTR. If I only allow the 3' UTR, the number of consistent ones should be higher. Another reason for the inconsistent positions might be, that the tools they used in miRTarBase use better and more specific alignment algorithms where more parameters can be defined. The tools can better be adapted to a biological background like seed region and additional 3' pairing. The pairwise2 module only has 4 scoring parameters to select and only delivers the alignment with the highest score and no ranking.

Table 3.2: Number of common alignment starting positions

	3 -2 -4 -4	5 -1 -8 -4	5 -2 -8 -3	5 -3 -8 -2	5 -4 -6 -4
Found number	815	804	388	411	800

	1 -2 -2 -1	2 -2 -5 -4	-4 -4	-5 -1
Found number	882	793	776	845

Chapter 4

Discussion

The main focus in this research was on sequence alignment of miRNA sequence and its target mRNA sequence for given validated miRNA target interactions. This alignment should simulate a binding between the two sequences. With this method, it was investigated how this approach can be used to predict new target sites for miRNAs without any other biological experiments.

The results showed that with this alignment tool, the increase in the alignment score for true targets is significant. This means that if a certain threshold for the scoring is exceeded, with a certain error rate a considered mRNA is truly targeted by this miRNA. However, the p-values are extremely low, which is very surprising. Observing the existing tools, they always take many features into account to make the prediction more reliable. Therefore, considering this one single feature can not be as significant as the p-values express.

The other part of the analysis, the analysis of the complementarities, did not show significant differences between true targets and non-targets. The ratio of complementary nucleotides of miRNA and mRNA of true targets were not significantly higher than the ones of the non-targets. This means, the alignment did not detect better binding sites for the true targets regarding complementarity. According to the p-value and the alignment scores, the alignment of the true targets was better and had more matches and less mismatches or gaps. This however, can not be observed from the plots very well. Considering only the complementarity without the alignment scores, no statement about the probability of a true target can be made. Additionally, the seed region and the additional 3' pairing, which should be present in binding sites, were not detected. Small increases in the first few nucleotides could be observed but not as clearly as expected. This may be due to the unspecific alignment tool. It finds the best alignment position concerning

the whole miRNA sequence. Therefore it does not set value on any special regions.

Another problem is that with this alignment tool different binding sites were found compared to binding sites predicted by some existing prediction tools like miRanda. Whether the positions I found with the alignment tool are the right ones is questionable. The alignment strongly depends on parameters which leads to the difference in results. Surprisingly, the parameter set with similar parameters to miRanda was one of the sets where the fewest consistent positions were found. As mentioned above one reason for the low number of consistent positions can be varying used databases which can differ from each other in terms of sequence sizes. A big disadvantage of the Biopython module is that it only delivers the best alignment and not for example the best three. Additionally, I considered the whole gene sequence as a target. If the best alignment was found in other regions than the 3' UTR, the starting position can not be found in the miRTarBase. In these cases, additional alignments with high scores would be helpful to analyse if for example the second best alignment score is the one that is also provided in the miRTarBase. Another limitation of the module is the small potential to adjust the parameters to the given problem.

Nevertheless the scores showed a significant increase for the true targets. Yet, it is questionable how reliable this prediction feature is if considered on its own. As can be observed for the existing tools, a much better prediction is achieved when considering more prediction features as described below.

A first step can be the improvement of the alignment process, especially the refinement of the alignment parameters. In some other tools, like miRanda, the parameters are adapted to the binding mechanism. For example, in the present seed region the complementarity should be nearly perfect, meaning mismatches in the first few nucleotides should be highly penalized or only a certain number of mismatches should be allowed. In addition, mismatches or gaps in the central nucleotides of the miRNA can be present and are therefore not much penalized. Adapting the parameters to some targeting rules can highly improve the target prediction. A more general improvement is that the more data of true targets is available, the more can be learned from the mechanism and the more the prediction tools can be adapted in different ways concerning features and precision of features.

The complexity of the prediction is highlighted when considering the consensus of all predicted targets of the different tools. Only a really small subset of all targets are predicted by all tools. Further, comparing validated target interactions of the database with the results of all tools, about 16% of all interactions are predicted by at least one tool. But there is no interac-

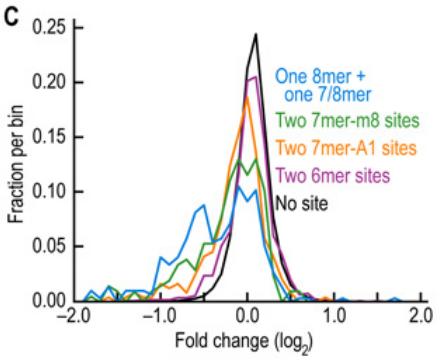

Figure 4.1: Effectiveness of different types of sites

tion that is predicted by all tools. That shows that there are big differences between the different tools partly because of the variable features they consider [Keller, 2016].

Because of their potential to function in treatment and detection of diseases, miRNAs become more and more important. Therefore miRNA research must be further extended and the complex target prediction has to be improved. Many researchers have found different additional features that can be considered. As described in the introduction, the seed region is one of the most important ones. The different types of sites vary in the effectiveness of repression. Grimson et al. (2007) investigate the different types and figure 4.1 shows their results. If one 8mer site is present, the repression effect is the highest [Grimson et al., 2007].

Another useful feature is the free energy of the binding complex. This energy determines the stability and tightness of a system, in this case the binding [Peterson et al., 2014]. If the free energy is lower, the binding between the miRNA and mRNA is tighter resulting in more evidence that the miRNA actually targets this mRNA. If the energy is high then the binding is not very favourable, meaning the miRNA does not favour to target this mRNA at this position. Some tools like RNAhybrid rely strongly on this feature [Rehmsmeier et al., 2004].

A completely different feature, the evolutionary conservation, can be taken into account. If a sequence occurs across species, it is defined as conserved. This implies that this part has been maintained by evolution because of a selected function [Peterson et al., 2014]. Conservation near the miRNA binding site can indicate that this part of the sequence is necessary for some mechanisms. This includes conservation of the miRNA itself as well as the conservation of the respective site of the mRNA. These conservations can be analysed with phylogenetic methods. Figure 4.4 shows how conserved the

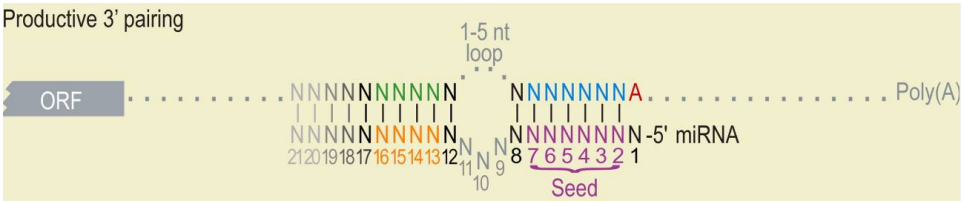

Figure 4.2: Additional 3' pairing

different nucleotides of a miRNA are, indicating regions like the seed region and additional 3' pairing.

Site accessibility is a feature that Kertesz et al. (2007) investigated for their target prediction. The mRNA generally folds into a secondary structure. Therefore the miRNA cannot easily bind to its target because at first, interactions within the mRNA have to be broken to make the target accessible. As a result, miRNA will favourably bind to regions where the mRNA is more accessible. The researchers found that if the targets form stem structures, the repression is reduced. If sites occur in open loop structures, the repression is much higher. Summing up they found that site accessibility is not less important than seed matching [Kertesz et al., 2007].

As already mentioned above, in addition to the seed matching towards the 5' end, another complementary site towards the 3' end in the miRNA is present (Figure 4.2). Grimson et al. (2007) investigated that the highest down regulation was found when the site started at position 13 and had four or five contiguous base pairings (Figure 4.3) [Grimson et al., 2007]. Considering again the conservation of the nucleotides they found that outside the seed region the contiguous nucleotides starting from nucleotide 12, 13 or 14 were the most conserved ones indicating their functional importance (Figure 4.4). Putting the information of seed region and additional base pairing together, it can be observed, if one 7mer-m8 site is present as well as a good 3' pairing, the efficacy is the highest. This is observed even though the difference between the efficacy of the presence of one 8mer site and the one mentioned before is not very big. But the improvement of the presence of a good 3' pairing instead of a poor one is more significant (Figure 4.5). Between the two complementary areas there is usually a part of non-pairing nucleotides where bulges and mismatches are found. These are important for the prevention of the AGO cleavage function [Filipowicz et al., 2008].

To get an even more reliable and precise prediction even more features can be considered. The presence of GU-wobbles is common in targeting. In this wobble positions Guanine(G) binds to Uracil(U) even though the pairing with Cytosine would be prevalent. This special pairing is thermo-

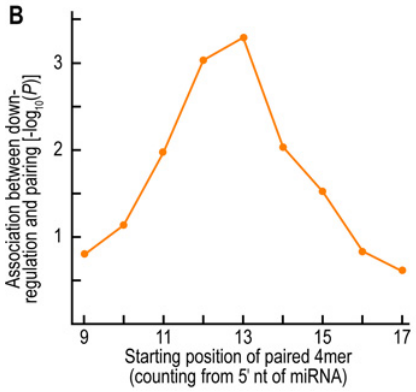

Figure 4.3: Relation between regulation and starting position of additional pairing

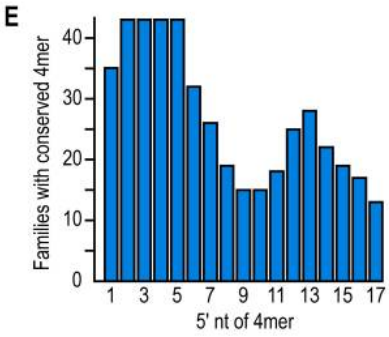

Figure 4.4: Conservation of nucleotide positions of miRNA

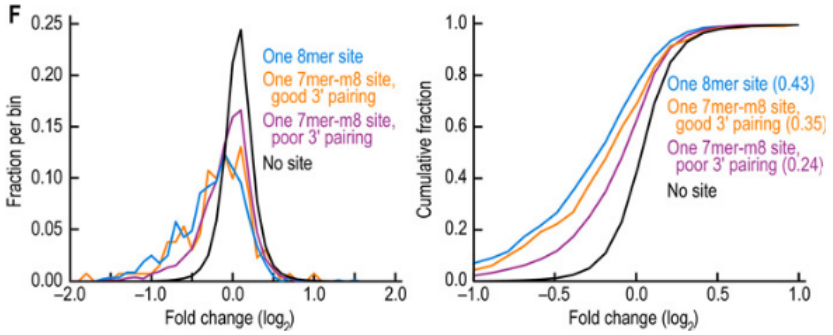

Figure 4.5: Efficacy of different combinations of seed region and additional 3' pairing

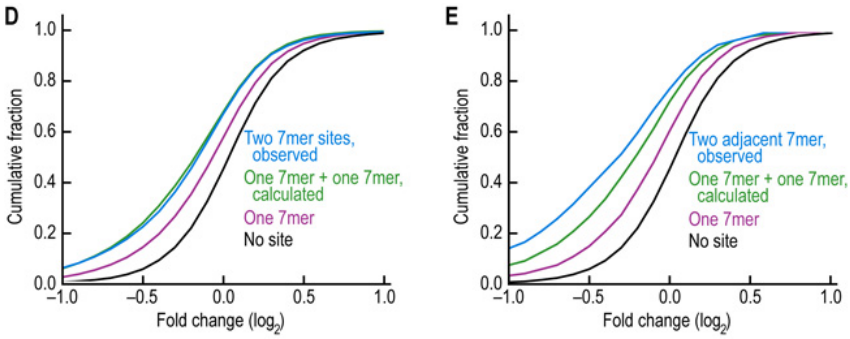

Figure 4.6: Effect of multiple sites

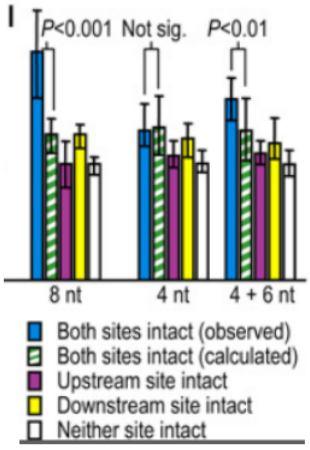

Figure 4.7: Cooperative repression with different site spacings

dynamically favourable. Therefore it occurs in many target interactions but also results in lower repression of the translation [Doench and Sharp, 2004]. As mentioned before, miRanda uses different scores for matches in the alignment step. They score the usual base pairing A-U and G-C with +5 and they penalize the other mismatches with a score of -3, excluding the pairing of G and U. This pairing is rewarded with at least +2 and therefore the GU wobbles are not penalized as much as other mismatches because they are very common [Enright et al., 2003].

Enright et al. (2003) and Doench et al. (2004) also found that the presence of multiple miRNA target sites results in a higher repression and destabilization of the mRNA. Grimson et al. (2007) further investigated that the distance between two sites is also an important criterion. Generally the repression of two present sites is the multiplication of the two single ones because they act independently. What is interesting is that if the two sites are adjacent, the repression is increased and not equal to the multiplication

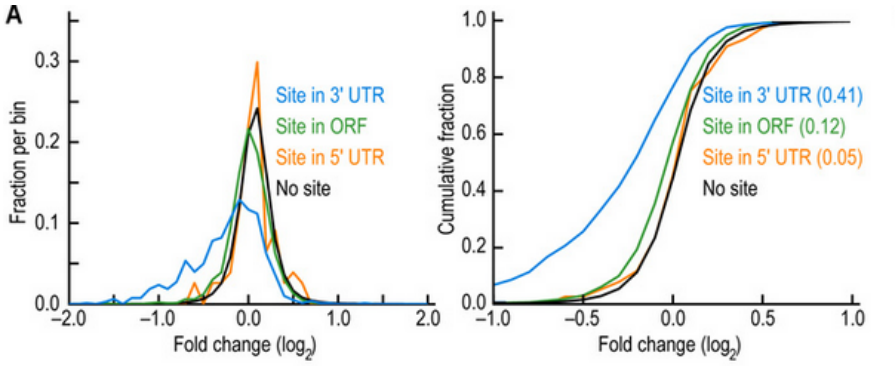

Figure 4.8: Efficacy of different site locations

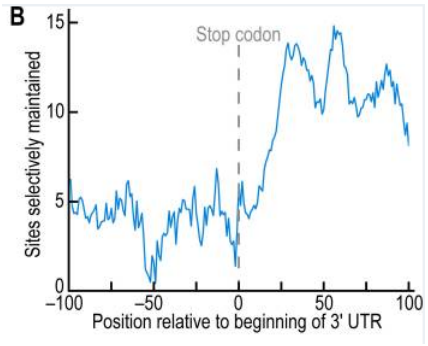

Figure 4.9: Efficacy of sites located relative to the stop codon

of the single ones. The increase in repression is however not very high (Figure 4.6). To investigate the effect of cooperative miRNAs they analysed a mixture of miR-1 and miR-133 and simulated three different spacings. The results show that a spacing of four nucleotides did not show a cooperative repression but six or eight nucleotide spacing showed an increase in repression (Figure 4.7).

Another indicator is the position of the binding site relative to the stop codon and the center of the UTR. Generally sites in the 3' UTR are investigated but Grimson et al. (2007) detected that sites in the Open reading frame (ORF) are slightly effective, sites in the 5' UTR not at all. Figure 4.8 shows the different efficacies regarding the site location. Another characteristic concerning the site locations is the distance from the stop codon. Figure 4.9 illustrates that approximately in the first 15 nucleotides the efficacy is still very low like in the ORF but afterwards it increases a lot. The sites were present at least 15 nucleotides from the stop codon and not present in the center of long UTRs.

FEATURES USED IN miRNA TARGET PREDICTION

Tool name	Seed match	Conservation	Free energy	Site accessibility	Target-site abundance	Machine learning	References
miRanda	X	X	X				Enright et al., 2003; John et al., 2004
miRanda-mirSVR	X	X	X	X		X	Betel et al., 2010
TargetScan	X	X					Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009; Garcia et al., 2011
DIANA-microT-CDS	X	X	X	X	X	X	Maragkakis et al., 2009; Reczko et al., 2012; Paraskevopoulou et al., 2013
MirTarget2	X	X	X	X		X	Wang, 2008; Wang and El Naqa, 2008
RNA22-GUI	X		X				Hofacker et al., 1994; Miranda et al., 2006; Loher and Rigoutsos, 2012
TargetMiner	X	X	X	X	X	X	Bandyopadhyay and Mitra, 2009
SVMicrO	X	X	X	X	X	X	Liu et al., 2010
PITA	X	X	X	X	X		Kertesz et al., 2007
RNAhybrid	X		X		X		Rehmsmeier et al., 2004; Kruger and Rehmsmeier, 2006

Figure 4.10: Prediction tools and their features

Considering all these different common and less common features, new targets for miRNAs can be predicted more reliably. Figure 4.10 shows known prediction tools and the features they consider. As mentioned above the most common feature that nearly all of them consider are seed matching, conservation and free energy.

Chapter 5

Conclusion

In this research validated miRNA target interactions from the database miRTarBase were investigated. With the help of sequence alignments I analysed potential binding sites. The main focus was on the alignment scores for these sites and whether they are higher for true targets then for non-targets. The results surprisingly showed that, according to the low p-values of a t-test, the consideration of the sequence alignment scores is already significant for the prediction of new targets. However, the analysis of the complementarities did not show big differences between true targets and non-targets. The significant seed region and additional 3' pairing could hardly be observed. Hence, the consideration of these features is questionable and still includes many errors. To improve the prediction many other features should be take into account. As described above, there are a lot of features that can be incorporated as well as many different tools that already exist. They all differ from each other in the number and type of features they use and also in the weighting of the single features [Peterson et al., 2014]. Concluding, each tool and feature has its advantages and strengths but also its limitations which makes none of the existing tools perfect and 100% reliable. Relying on just a few features will for sure lead to many errors, either false positives or false negatives. There are many irregularities and some features like the free energy are not always very precise [Peterson et al., 2014]. These problems have to be incorporated to lower the error rate.

Additionally not every tool is frequently updated which is a big problem and does not lead to improvements in target prediction. Some tools do not use the current data or innovations in target interactions. According to Peterson et al. (2014) the following tools are outstanding because of maintenance, newest input data and they are the easiest ones to use: DIANA-microT-CDS, miRanda-mirSVR and Targetscan [Peterson et al., 2014]. All three are somehow unique but all for example use looser thresholds for conservation allowing less conserved regions to balance irregularities and refuse

less true targets.

The early miRanda tool is also still widely used. Its strengths are the optimization of the sequence complementarity. This is due to the specific rules and requirements for example regarding the sequeunce alignment and conservation. However, the tool permits imperfect bindings. Additionally, it emphasizes the presence of multiple binding sites in the 3' UTR, either of the same miRNA or of others. This might result in refusing miRNAs with only one site leading to many false positives [John et al., 2004]. In contrast, TargetScan places emphasis on the seed region with its many different seed types and context scores. Having binding sites with less perfect seed regions even though they are functional, the tool would ignore them [Lewis et al., 2005]. The last favoured tool, DIANA-microT-CDS, predicts only single binding sites, which can be conserved and non-conserved [Kiriakidou et al., 2004]. Certainly, the presence of multiple binding sizes might result in weaker interactions that are therefore underestimated by this tool. However, with its signal to noise ratio, which evaluates the specificity, and the resulting prediction score, which correlates with the amount of produced protein, DIANA-microT-CDS is an innovative, successful prediction tool [Maragkakis et al., 2009]. The class of tools, like miRSVR, learn how to weight the different features they consider with support vector regression [Betel et al., 2010]. These machine-learning based approaches are getting more accurate the more positive and negative target data is verified and the more features are identified. Consequently, because of the lack of this data, these tools are not significantly better than other tools mentioned above [Peterson et al., 2014].

Fujiwara and Yada (2013) followed a new approach by considering other characteristics than binding sites for the prediction, such as the transcriptional regulation. In their research they searched for common cis-elements in the miRNA as well as in the target gene. Compared to conventional methods, their method is almost as good as the standard binding site based ones but combining the two different methods decreases the accuracy a lot. The advantages of this novel approach are independence of conservation of binding sites and the amount of available training data [Fujiwara and Yada, 2013].

Coronnello and Benos (2013) investigated another approach trying to improve the prediction power. They developed a tool, ComiR, that additionally considers the miRNA expression levels and the combination of miRNA bindings. It also combines different scoring schemes from tools mentioned above. The innovation in this tool is the investigation of sets of miRNAs and their co-expression [Coronnello and Benos, 2013].

All in all, the analysis of miRNAs is very complex and needs to be improved in terms of prediction. The future consists of elimination of the limits of some tools and finding more useful features for the prediction.

Bibliography

[Ardekani and Naeini, 2010] Ardekani, A. M. and Naeini, M. M. (2010). The Role of MicroRNAs in Human Diseases. *Avicenna J Med Biotechnol*, 2(4):161–179.

[Bartel, 2004] Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297.

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

[Bioinformatics.org, 2010] Bioinformatics.org (2010). Definition of sequence alignment. URL `http://www.bioinformatics.org/wiki/sequence_alignment`.

[Biopython, 2016] Biopython (2016). Documentation of module pairwise2. URL `http://biopython.org/DIST/docs/api/Bio.pairwise2-module.html`.

[Chou et al., 2016] Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. D., Lin, Y. L., Lee, W. H., Yang, C. D., Hong, H. C., Wei, T. Y., Tu, S. J., Tsai, T. R., Ho, S. Y., Jian, T. Y., Wu, H. Y., Chen, P. R., Lin, N. C., Huang, H. T., Yang, T. L., Pai, C. Y., Tai, C. S., Chen, W. L., Huang, C. Y., Liu, C. C., Weng, S. L., Liao, K. W., Hsu, W. L., and Huang, H. D. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, 44(D1):D239–247.

[Coronnello and Benos, 2013] Coronnello, C. and Benos, P. V. (2013). ComiR: Combinatorial microRNA target prediction tool. *Nucleic Acids Res.*, 41(Web Server issue):W159–164.

[Doench and Sharp, 2004] Doench, J. G. and Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. *Genes Dev.*, 18(5):504–511.

[Enright et al., 2003] Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biol.*, 5(1):R1.

[Filipowicz et al., 2008] Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, 9(2):102–114.

[Friedman et al., 2009] Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105.

[Fujiwara and Yada, 2013] Fujiwara, T. and Yada, T. (2013). miRNA-target prediction based on transcriptional regulation. *BMC Genomics*, 14 Suppl 2:S3.

[Griffiths-Jones et al., 2006] Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, 34(Database issue):D140–144.

[Grimson et al., 2007] Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, 27(1):91–105.

[Ha and Kim, 2014] Ha, M. and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, 15(8):509–524.

[Hsu et al., 2011] Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., Tsai, W. T., Chen, G. Z., Lee, C. J., Chiu, C. M., Chien, C. H., Wu, M. C., Huang, C. Y., Tsou, A. P., and Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, 39(Database issue):D163–169. URL `http://mirtarbase.mbc.nctu.edu.tw/`.

[John et al., 2004] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA targets. *PLoS Biol.*, 2(11):e363.

[Karolchik et al., 2004] Karolchik, D., Hinrichs, A., Furey, T., Roskin, K., Sugnet, C., Haussler, D., and Kent, W. (2004). The ucsc table browser data retrieval tool. URl `https://genome.ucsc.edu/`.

[Keller, 2016] Keller, A. (2016). Special lecture: Algorithms and applications for microrna research.

[Kertesz et al., 2007] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.*, 39(10):1278–1284.

[King, 2013] King, D. (2013). Documentation/how tos/calc: Ttest function. URL `https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_TTEST_function`.

[Kiriakidou et al., 2004] Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, 18(10):1165–1178.

[Kozomara and Griffiths-Jones, 2011] Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39(Database issue):D152–157. URL `http://mirbase.org/`.

[Lee et al., 2002] Lee, Y., Jeon, K., Lee, J. T., Kim, S., and Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, 21(17):4663–4670.

[Lewis et al., 2005] Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

[Macfarlane and Murphy, 2010] Macfarlane, L. A. and Murphy, P. R. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. *Curr. Genomics*, 11(7):537–561.

[Maragkakis et al., 2009] Maragkakis, M., Alexiou, P., Papadopoulos, G. L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V. A., Sethupathy, P., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., and Hatzigeorgiou, A. G. (2009). Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, 10:295.

[Mudunuri et al., 2009] Mudunuri, U., Che, A., Yi, M., and Stephens, R. (2009). biodbnet: the biological database network. URL `https://biodbnet-abcc.ncifcrf.gov/db/db2db.php`.

[Peterson et al., 2014] Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L., and Congdon, C. B. (2014). Common features of microRNA target prediction tools. *Front Genet*, 5:23.

[Rehmsmeier et al., 2004] Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517.

[Rhoades et al., 2002] Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, 110(4):513–520.

List of Figures

List of Tables