Contents

1 Introduction

1.1 Background

miRNAs are small, non-coding RNAs that function in post-transcriptional regulation of gene expression, especially in terms of gene silencing. A miRNA consists of approximately 22 nucleotides which build a single strand RNA and are processed by Drosha and Dicer. After the processing the initial double stranded miRNA-duplex the guide strand will be further used in association with other molecules [4]. The miRNA is therefore mainly active in combination with a protein of the AGO protein family and with this it builds an RNA-induced silencing complex [13]. This complex targets a mRNA mainly at its 3' untranslated region (UTR) by complementary binding to the sequence. This binding results in gene silencing either by repression of mRNA translation or degradation of the respective mRNA [10].

This gene regulation by miRNAs plays an important role for many major cell functions like growth, differentiation or metabolism [5] and is currently examined by many investigators. The importance can also be observed by the fact that more than 60% of all human genes have a conserved miRNA-binding site [13]. Therefore disregulation, either up- or downregulation, may result in different human diseases, especially cancer [5]. This also leads to the assumption that miRNAs are a useful feature for the diagnosis and treatment of diseases. But the first step is to understand their function and regulation mechanism. With these information it can be possible to predict targets of miRNAs and then specifically effect these target interactions.

But especially this prediction of new targets of miRNAs is the main challenge in the whole field of miRNAs. There is still no perfect, reliable solution for it. There are a few tools that try to solve this problem. One of the first tools was miRanda. In their prediction algorithm they rely on three main features: sequence complementary, free energy and evolutionary conservation [10]. Another tool is miRSVR that uses bit vectors and statistical learning approaches to predict targets. In this bit vector they store information like base complementarity, UTR length and distance, AU content and conservation [6]. A third tool was developed called TargetScan. Their approach is based on seed matching and additional features like UTR positioning, AU content as well and base pairing of the 13nt to the 16nt miRNA nucleotide. They especially take the seed region into account by defining four different patterns of seed binding sites [19].

The challenge of the prediction can be observed by considering the consensus of all predicted targets of the different tools. Just a really small subset of all targets are predicted by all tools. If you then compare validated target interactions with the results of all tools about 16% of all interactions are predicted by at least one tool. But there is no interaction 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. (Quelle Keller Vorlesung 4)

Because the miRNA binds to the sequence of its target gene an obvious starting point can be the complementarity of the miRNA sequence to the gene sequence as almost each tool considers at first. The first thought could be that if the sequence complementarity is quite high this could be interpreted as a new target site. Considering the sequence alignment a high alignment score of miRNA and mRNA sequence can indicate a true target. Whether this is a reliable indication or not will be discussed below. For this assumption there already exist some problems beforehand. Because it is known that the main binding happens within the seed region, other regions are not necessarily complementary and this may lower the complementarity or alignmnet score.

Because of the early knowledge of base complementarity close to the 5'end of the miRNA, one of the first features that was used for the prediction, is the presence of the so called 'seed region' of the miRNA. This region consists of 7 or 8 nucleotides that are complementary to the respective nucleotides in the mRNA sequence. Another early characteristic is the bulge after the complementary region where the nucleotides do not fit together. After this part complementary bases towards the 3' end of the miRNA may occur again. Whether these rules apply and if they are present in validated data will be discussed later.

For these miRNA analyses three main databases are useful and necessary. A collection of experimental validated miRNA-target interactions(MTIs) is essential. The database miRTarBase provides a big dataset of validated MTIs [14]. To get information about every single miRNA for different organisms miRBase contains large datasets about the miRNAs. The most interesting information are sequences of the precursor and of all mature miRNAs [17]. For further analyses I will only consider miRNAs of *Homo sapiens*. The last used database is the UCSC Genome Bioinformatics Site. In their table browser you can get data files about human gene sequences and untranslated regions and their respective sequences [15].
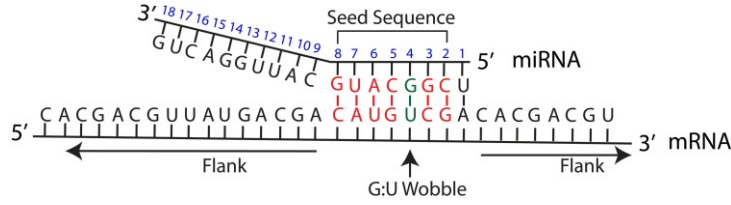
Figure 1.1: Scheme of seed matching

1.2 Prediction features

1.2.1 Seed matching

The main feature that is considered in this research 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. There can be regions were the nucleotides are unbound which results in complex secondary structures that are hard to predict [22] (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1370637/) The main complementary region or seed region includes nucleotides from position two to eight starting from the 5' end. Figure 1.1 shows the scheme of the seed region (Peterson et al., 2014). This irregularity of the presence of non-complementarity makes the reliable prediction more difficult for animals than for plants and only concentrating on the seed region will lead to many false positives. How significant the consideration of the sequence complementarity of miRNA and mRNA sequeence is, will be further analysed in this research.

In the mentioned seed region there can be different types of matching patterns: perfect match between six nucleotides, seven nucleotides including the 8th position, seven nucleotides and an Adenine at position one to eight nucleotides and also an Adenine at position one [19] [7] [18]. Figure 1.2 illustrates the different types of sites. (http://www.targetscan.org/docs/7mer.html) Grimson et al. [12] investigated the distribution of the different types. Figure 1.3 shows that the repression is the highest if an 8mer site is present.

1.2.2 Free energy

The free energy of a system determines its stability and tightness of binding. If the free energy is lower than the binding between the miRNA and the
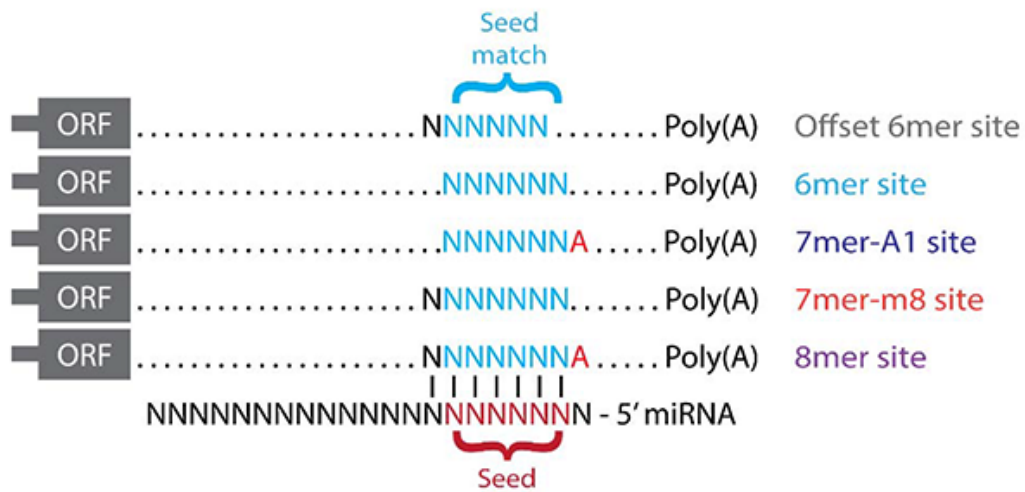
Canonical site types
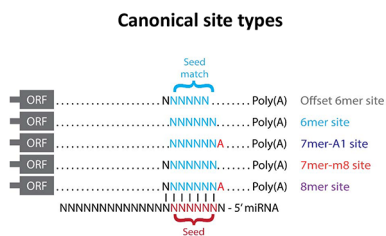

Figure 1.2: Canonical sites of seed region

Figure 1.3: Effectiveness of different sites

mRNA sequence is tighter resulting in more evidence that this miRNA targets this mRNA. If the energy is high then the binding is not very favourable meaning the miRNA not favours to target this mRNA at this position. Some tools like RNAhybrid [22] rely mainly on this feature. (thermodynamics von miranda)

1.2.3   Conservation

Another feature to take into account is the evolutionary conservation. If a sequence occurs across species it is defined as conserved. This implicates that this part has been maintained by evolution because of a selected function [21]. Conservation near the miRNA binding site can indicate that this part of the sequence is necessary for some mechanisms. This includes convercation of the miRNA itself as well as the conservation of the respective site of the mRNA. These conservations can be analysed with phylogentic methods.

1.2.4   Site accessibility

Kertesz et al. [16] investigated the importance of site accessibility for the target prediction. (http://www.nature.com/ng/journal/v39/n10/full/ng2135.html) The mRNA generally folds into a secondary structure. Therefore the miRNA can not 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. Kertesz et al. [16] found that if the targets form higly stem structures the repression is reduced. If sites occur in open loop structures the repression is much higher. Summing up they found that site accessibilty is not less important than seed matching.

Quelle: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3927079/

1.2.5   Addtional Watson-Crick pairing

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3800283/ Grimson et al. 2013 In addition to the seed matching towards the 5' end, another complementary site towards the 3' end in the miRNA is present. Grimson et al. [12] investigated that the highest down regulation was found when the site started at position 13 and had four or five contiguous base pairings (Figure 1.2.5).

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 / demonstrating their functional importance (Figure 1.2.5. Putting the information of seed region and additional base pairing together it can be observed that if one 7mer-m8 site is present as well as a good 3' pairing that the efficacy is the highest. 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.

In between the two complementary areas there is usually a part of non pairing nucleotides where bulges and mismatches are found.

1.2.6 Other features

To get an even more reliable and precise prediction additional 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 thermodynamically favourable and occurs therefore in many target interactions but results in lower repression of the translation. (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC374233/) As mentioned before miRanda uses different scores for matches in their 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 [10].

https://elifesciences.org/content/4/e05005

Enright et al. [10] and Doench et al. [9] (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC37423 also found that the presence of multiple miRNA target sites results in a higher repression and destabilization of the mRNA. Grimson et al. [12] (http://www.cell.com/molecular-cell/fulltext/S1097-2765(07)00407-8) further investigated that the distance between two sites is also an important criterion. Generally the repression of two present site is the multiplication of the two single once because they act independently. The interesting thing now is that if the two sites are adjacent the repression is increased and not equal to the multiplication of the single ones. The increase in repression is however not very high (Figure 1.5) 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 4 nt did not show a cooperative repression but 6 or 8 nt spacing showed an increase in repression (Figure 1.4).

Another indicator is the position of the binding site relatively to the stop codon and the center of the UTR. Generally sites in the 3'UTR are investigated but Grimson et al. detected that sites in the Open reading frame (ORF) are slightly effective, sites in the 5'UTR not at all [12]. Figure 1.6 shows the different efficacies. Another characteristic concerning the site locations it the distance from the stop codon. Figure 1.7 illustrates that in the first 15 nt the efficacy is still very low like in the ORF but afterwards it increases much. The sites were present at least 15 nucleotides from the stop codon and not present in the center of long UTRs but away from it. Grimson et al. also found that the AU nucleotide content is increased in the region of conserved sites [12].

Figure 1.8 shows known prediction tools and the features they consider. As mentioned above the most common feature that early all of them consider
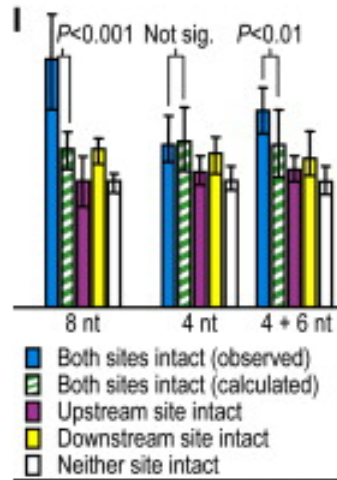
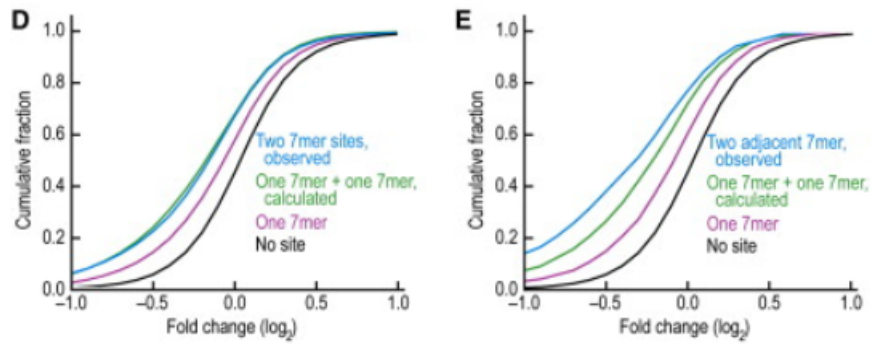Figure 1.4: Cooperative repression with different site spacings
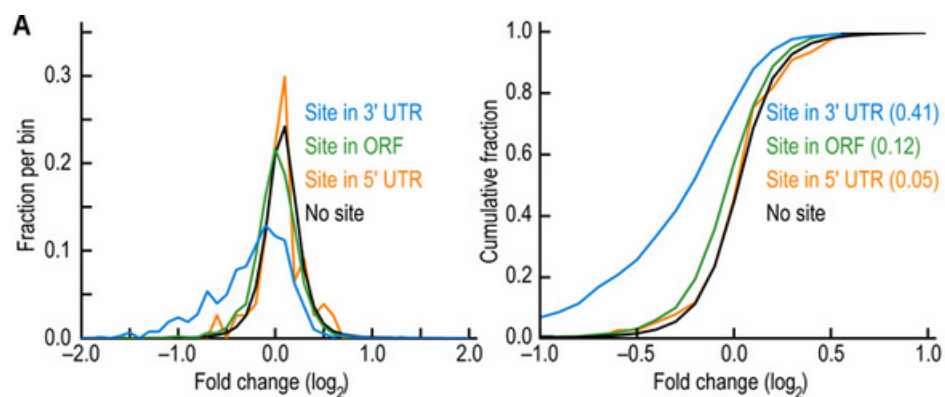

Figure 1.5: Effect of multiple sites
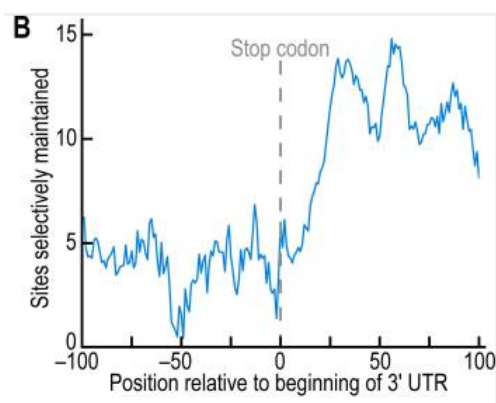
Figure 1.6: Efficacy of different site locations

Figure 1.7: Efficacy of sites located relative to the stop codon

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 1.8: Prediction tools and their features

are seed matching, conservation and free energy.

2 Methods and Data

2.1 Methods

The program for the analysis is implemented in Python. The data files are parsed and stored into dictionaries that allow fast searching. Then for every MTI in the miRTarBase file the respective miRNA sequence is searched in the miRBase file and combined with the gene sequence of the corresponding target in the UCSC file. Having theses two sequences, a local alignment is performed. 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. From the result functional or structural similarities can be obtained [1]. In this case, the similarity can be interpreted as a binding site, because 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 alignment is executed, not a global.

The Biopython library provides a module, pairwise2, for pairwise local

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

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 [2]. To get a more suitable alignment own parameters can be selected. Table 2.1 shows the different parameters that I used to generate the data. Set no. 8 is similar to the parameters they used for the tool miRanda [10]. The other parameters are just logically selected to test which influence they have on the results.

To be able to analyse the data statistically a set of negative controls is required. To produce this data, miRNAs were randomly assigned to genes and then aligned in the same way as the sequences of the true target interactions. For these datasets the average and standard deviation were computed as well. Then with case and control alignment scores a statistical two-tailed t-test for two samples with equal variances was performed with LibreOffice [3]. This provides a p-value as a result. The lower this p-value is, the more significant the increase of the alignment score of the true targets is. If there is an increase in the score, this would be an indication that is we only calculate the alignment score for a new miRNA and any target sequence and observe a high one, the probability that this is a true target would be high. If this is a reliable feature for the prediction will be discussed in the following.

2.2 Data

2.2.1 miRTarBase

The database mirTarBase which was released in 2010 provides by now about 7500 strong validated MTIs and 348000 weak ones from different species [8]. In this research I concentrate on humans. Different experiment types were used to validate the data, including Reporter assay, Western plot, qPCR, Microarray, NGS, pSILAC and other methods where the first three are the ones that deliver strong evidences [14]. Again in here, I only concentrate on the strong evidence targets. In detail, the data collection provides many 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 you can directly download the respective MTI data tables of the favoured species, in this case *Homo sapiens*. The table then contains the following information: 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 interesting fields for the research are name of the miRNA, the target gene ID and the experiment type because I only concentrate on the Functional MTIs that are not weak. In fact, every single strong validated interaction was analysed.

As mentioned abovem, the miRTarBase provides also binding sites as alignment positions predicted by different tools. To decide whether my own found alignments are compatible with the provided ones, the miRTarBase html page was parsed to get the start positions of each provided alignment. These positions exist not 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.

2.2.2 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 [11]. By now it contains about 35000 mature forms and 2500 of it 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 of the IDs and the sequences to align those to the gene sequence [17].

2.2.3   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 Bioinformatics Site you can generate a list of all genes and their UTRs of the human genome using the Table Browser [15]. The 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.

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. But 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 [20]. I entered all existing target Entrez IDs of the miRTarBase file and obtained a list of IDs and their corresponding Refseq accession numbers. In my program I used this list by storing every entry in a dictionary and simply looked the particular ID up for every MTI.

3   Results

For each parameter set the alignment score for every MTI is stored in a table. Additionally the average alignment score, the standard deviation and the p-values are listed underneath. Table 3 shows only a summary of the table, excluding the single alignment scores. (whole table somewhere else) 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 slightly higher than the ones of the non targets. For the lower alignment scores of e.g. set 1 -2 -2 -1, resulting from really low match scores, the difference is only 0.3, so not significant. For the higher scores the difference amounts to 2. The standard deviation of the non targets is slightly higher than for the true targets but also not very significant. The p-value sheds light on whether the increase in

Table 3.1: Table of alignment results

	3 -2 -4 -4	5 -1 -8 -4	5 -2 -8 -3	5 -3 -8 -2	5 -4 -6 -4
Average true targets	33.096	64.247	59.495	56.391	55.510
Average non targets	31.831	62.199	57.464	54.269	53.294
Standard deviation true targets	4.058	5.758	6.304	6.633	6.806
Standard deviation non targets	4.233	6.510	6.753	6.876	7.01
t-test p-value	5.934E-49	2.529E-61	1.040E-51	1.394E-51	1.744E-53

	-4 -4	-5 -1	1 -2 -2 -1	2 -2 -5 -4
Average true targets	13.712	13.710	8.593	18.351
Average non targets	13.362	13.361	8.208	17.619
Standard deviation true targets	1.078	1.078	1.347	2.751
Standard deviation non targets	1.311	1.309	1.385	2.868
t-test p-value	8.824E-50	1.131E-49	8.157E-42	3.279E-36

the alignment score is significant when considering the whole set of scores. According to the listed values of the t-test this is a significant increase in the score because they are all really low. The parameters 5 -1 -8 -4 show the most significant difference with a p-value of 2.529E-61 whereas 2 -2 -5 -4 shows the least significant increase but also here the p-value is really low with 3.279E-36. So even though the average scores are not significantly different, the distribution of the score seem to be higher in the true targets than in the non targets.

This increase in the alignment score can be also be observed when considering the distribution of the scores within the two groups of case and control. For two parameter sets such a distribution was plotted which is shown in Figure 3.2. On the x-axis the different scores are shown and on the y-axis the percentage of how many of the true targets had these scores. It can be observed that towards the higher alignment scores the percentage of true targets with this score increases. Although there are some scores that are exceptions because more non targets produced this scores than true targets. Assuming the score as the only prediction feature, the high score would classify this target as a true target. This causes therefore still many false positives. But in general the higher the score the higher the likelihood to observe a true target.
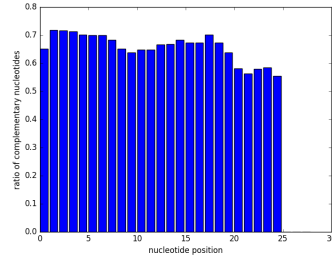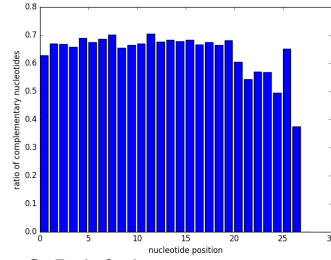
For every parameter set the complementarity of the alignments of the miRNA and gene sequence were also plotted. These are shown in Figure **??** to Figure **??**.
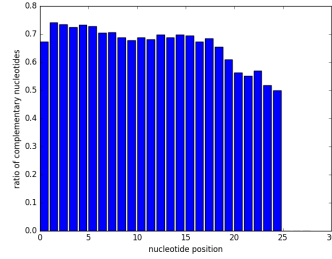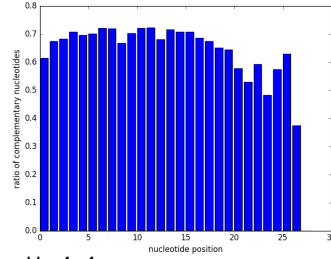
14
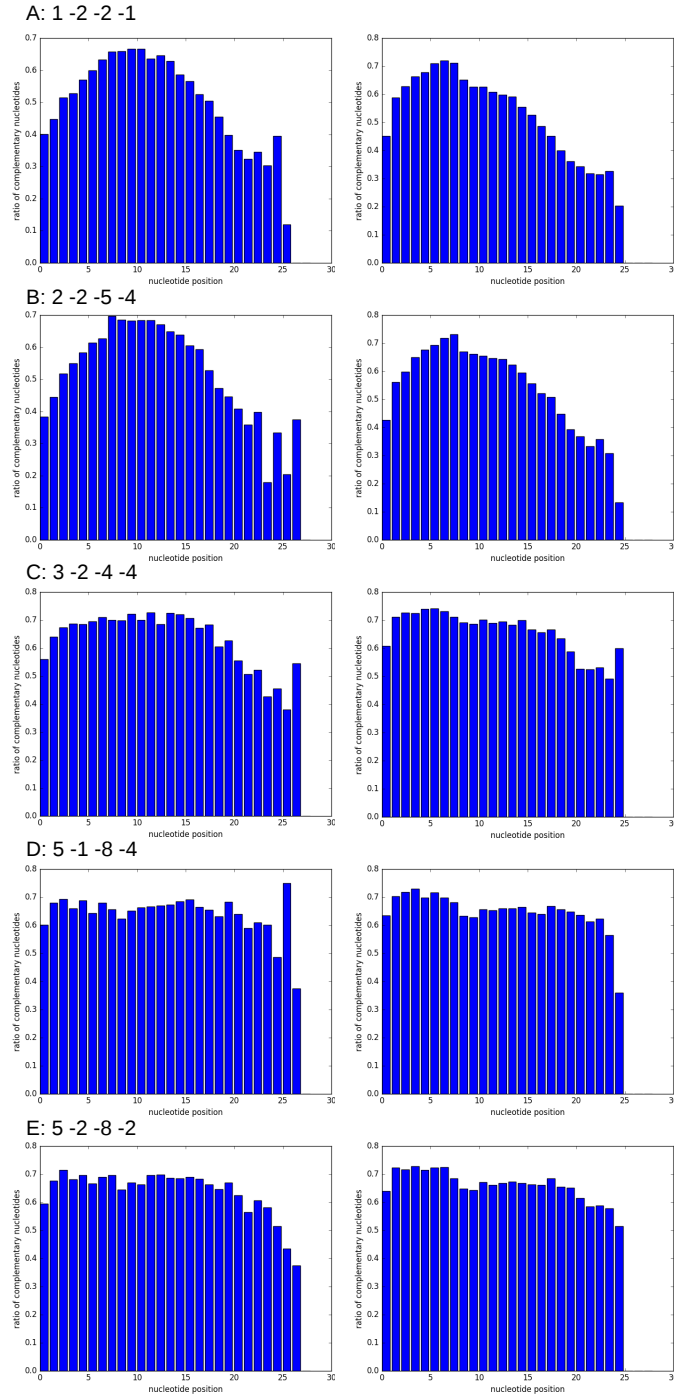
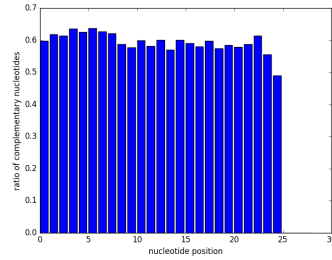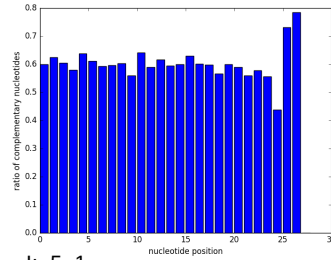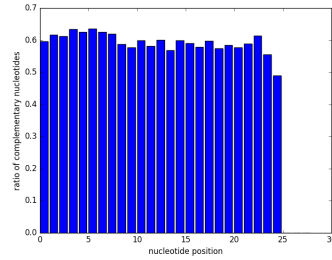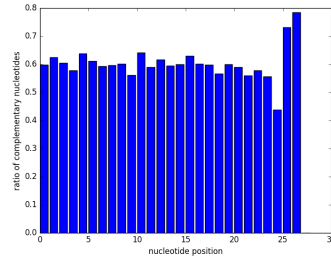Figure 3.1: Ratios of complementarity per parameter set: targets right, non targets left

F: 5 -3 -8 -2
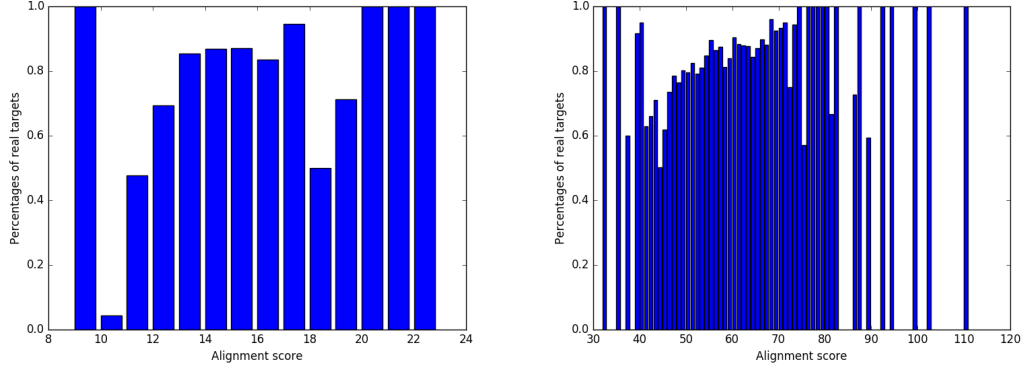
G: 5 -4 -6 -4

H: -4 -4

I: -5 -1

Figure 3.2: Distribution of alignment scores, parameters -4 -4 and 5 -3 -8- 2

The x-axis of the plots shows the positions one to 22 or more of the miRNA, the y-axis the ratio of complementary bases at this position considering all alignments produced with these parameters. For some sets there is not a significant difference between true targets and non targets, e.g. for parameters -5 -1 . Regarding set 1 -2 -2 -1 and for the true target an increase in complementarity in the seed region can be observed. On the other hand the ratio towards the end gets really low down to only 30% of the bases were complementary to the target sequence. In comparison, maximum of complementarity of the non targets is shifted towards the central positions, not showing the typical seed region. Notably the plot of the true targets does not show the complementary region towards the 3' end. Considering the other figures of the complementarities there are no significant differences between the two groups. For the non targets the complementarity is almost as high as for the true targets, gaining no reliable information for the prediction. This implicates that the consideration of the complementarities of the miRNA positions is not a significant and reliable prediction feature.

The analysis of the alignment starting positions was not really successful and significant. Table 3 shows the number 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. These two sets also use very similar parameters. This shows that the alignment positions strongly depend on the selected parameters. In comparison to the given number of about 3700 provided positions in the miRTarBase the numbers of the respective own found ones are in general not very high. This can be due to the different sizes of the

17

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

UTRs of the genes. In this research they are from the UCSC website whereas the miRTarBase may use another source. Therefore the sizes can be a bit different. To eliminate these small disagreements 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 if the position lies in a window of +5 or -5 related to the miRTarBase position.

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
0	0.598	0.597	0.452	0.427	0.608	0.634	0.641	0.652	0.673
1	0.619	0.618	0.588	0.562	0.712	0.703	0.724	0.718	0.741
2	0.614	0.613	0.628	0.599	0.727	0.718	0.717	0.716	0.735
3	0.635	0.634	0.664	0.650	0.725	0.730	0.729	0.713	0.724
4	0.626	0.625	0.678	0.676	0.740	0.698	0.716	0.701	0.734
5	0.637	0.636	0.710	0.694	0.742	0.716	0.723	0.701	0.729
6	0.627	0.625	0.720	0.718	0.732	0.699	0.724	0.700	0.705
7	0.621	0.620	0.712	0.731	0.712	0.682	0.685	0.683	0.707
8	0.588	0.588	0.651	0.669	0.692	0.633	0.649	0.652	0.689
9	0.577	0.577	0.626	0.662	0.687	0.628	0.643	0.638	0.679
10	0.599	0.599	0.626	0.655	0.701	0.657	0.672	0.649	0.689
11	0.582	0.581	0.608	0.647	0.690	0.654	0.662	0.648	0.681
12	0.601	0.601	0.598	0.643	0.696	0.659	0.668	0.666	0.698
13	0.570	0.569	0.591	0.624	0.683	0.660	0.673	0.669	0.688
14	0.600	0.600	0.555	0.595	0.700	0.666	0.669	0.683	0.698
15	0.590	0.590	0.527	0.557	0.667	0.645	0.664	0.673	0.695
16	0.580	0.579	0.487	0.522	0.657	0.641	0.661	0.673	0.673
17	0.599	0.598	0.452	0.508	0.666	0.668	0.684	0.702	0.685
18	0.574	0.575	0.400	0.448	0.635	0.656	0.655	0.673	0.655
19	0.585	0.585	0.361	0.393	0.588	0.649	0.652	0.638	0.611
20	0.578	0.577	0.344	0.369	0.527	0.637	0.615	0.582	0.563
Average	0.600	0.599	0.570	0.588	0.680	0.668	0.677	0.673	0.688
Average 1 – 8	0.621	0.620	0.669	0.662	0.723	0.697	0.708	0.698	0.720
Average 12 - 16	0.585	0.585	0.540	0.575	0.677	0.653	0.667	0.675	0.688

Negative control	-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
0	0.599	0.599	0.401	0.383	0.560	0.601	0.595	0.628	0.615
1	0.625	0.624	0.447	0.445	0.639	0.679	0.676	0.671	0.676
2	0.605	0.604	0.515	0.518	0.674	0.693	0.715	0.668	0.683
3	0.579	0.579	0.529	0.549	0.686	0.660	0.681	0.658	0.709
4	0.638	0.638	0.570	0.583	0.685	0.688	0.696	0.689	0.697
5	0.612	0.612	0.599	0.614	0.695	0.643	0.666	0.675	0.701
6	0.593	0.593	0.633	0.627	0.710	0.680	0.690	0.687	0.722
7	0.597	0.597	0.658	0.697	0.701	0.657	0.697	0.701	0.721
8	0.603	0.602	0.660	0.685	0.698	0.624	0.646	0.654	0.669
9	0.561	0.561	0.667	0.682	0.722	0.651	0.671	0.665	0.703
10	0.642	0.642	0.667	0.683	0.700	0.663	0.663	0.671	0.722
11	0.590	0.591	0.637	0.683	0.727	0.667	0.697	0.705	0.723
12	0.617	0.617	0.646	0.671	0.685	0.670	0.698	0.677	0.681
13	0.595	0.595	0.629	0.649	0.726	0.674	0.687	0.684	0.716
14	0.600	0.600	0.587	0.639	0.720	0.684	0.684	0.678	0.709
15	0.631	0.631	0.565	0.606	0.707	0.692	0.691	0.684	0.709
16	0.601	0.601	0.525	0.594	0.671	0.665	0.684	0.666	0.687
17	0.599	0.598	0.505	0.528	0.683	0.654	0.663	0.674	0.676
18	0.567	0.566	0.455	0.472	0.605	0.632	0.647	0.665	0.652
19	0.600	0.599	0.398	0.446	0.627	0.683	0.669	0.682	0.645
20	0.589	0.590	0.352	0.408	0.554	0.639	0.624	0.605	0.579
Average	0.602	0.602	0.554	0.579	0.675	0.662	0.673	0.671	0.685
Average 1 – 8	0.602	0.606	0.576	0.590	0.686	0.666	0.683	0.676	0.697
Average 12 - 16	0.601	0.607	0.576	0.622	0.706	0.679	0.686	0.678	0.705

Figure 3.3: Table of ratios of complementarities

References

[1] ?? Definition of sequence alignment. URL `https://en.wikipedia.org/wiki/Sequence_alignment`.

[2] ?? Documentation of module pairwise2. URL `http://biopython.org/DIST/docs/api/Bio.pairwise2-module.html`.

[3] ?? Documentation of ttest function of libreoffice. URL `https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_TTEST_function`.

[4] F. Afonso-Grunz and S. Muller. Principles of miRNA-mRNA interactions: beyond sequence complementarity. *Cell. Mol. Life Sci.*, 72(16):3127–3141, Aug 2015.

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

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

[7] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microRNA-target recognition. *PLoS Biol.*, 3(3):e85, Mar 2005.

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

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

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

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

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

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

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

[15] D. Karolchik, A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, and W.J. Kent. The ucsc table browser data retrieval tool, Jun 2004. URl https://genome.ucsc.edu/.

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

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

[18] A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microRNA target predictions. *Nat. Genet.*, 37(5):495–500, May 2005.

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

21

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

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

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

4 Abbildungen

Figure 1.1: http://journal.frontiersin.org/article/10.3389/fgene.2014.00023/full
Figure 1.2: http://www.targetscan.org/docs/7mer.html
Figure sites: http://www.ncbi.nlm.nih.gov/pubmed/17612493
Table 1.8: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3927079/