Saarland University
Center for Bioinformatics
Chair for Clinical Bioinformatics

Bachelor's Thesis

MiRNA target analysis focussing on sequence alignment

submitted by

Louisa Schwed

on September XX, 2016

Supervisor
Advisor I

Advisors
XXXX

Reviewers
Second Reviewer

Center for Bioinformatics

Schwed, Louisa
MiRNA target analysis focussing on sequence alignment
Bachelor's Thesis in Bioinformatics
Saarland University Saarbrücken, Germany
September 2016

Declaration

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

Saarbrücken, September xx, 2016

Louisa Schwed

Contents

Abstract

blablabla

1 Introduction

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. The maturation of the miRNA is shown in figure 1.1. After transcription the miRNA folds into a hairpin structure and builds the primary miRNA [Macfarlane and Murphy, 2010]. First, this pri-miRNA is processed by Drosha which results in a precursor miRNA (pre-miRNA). This precursor is about 60 to 70 nucleotides long and contains two single stranded miRNAs. After the export into the cytoplasm another processing step is executed with Dicer to gain the actual form of miRNA. But they are still double stranded and need to be unwinded to be mature miRNAs [Macfarlane and Murphy, 2010]. So after processing of the initial double stranded miRNA-duplex, the guide strand will be further used in association with other molecules [Afonso-Grunz and Muller, 2015]. In detail, the miRNA is mainly active in combination with a catalytic protein of the AGO protein family. With it, 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. The miRNA functions in this case as a guide with their base pairing [Macfarlane and Murphy, 2010]. It can also bind at the 5' UTR or even in the coding region but these are the uncommon regions. Either way, this binding results in gene silencing either by repression of mRNA translation or degradation of the respective mRNA [Enright et al., 2003]. Which of the two pathways is chosen depends on the degree of complementarity. As figure 1.1 illustrates, if the complementarity is nearly perfect then the target mRNA is degraded otherwise if the seed regions contains mismatches the translation is only repressed. Both ways have a huge impact on the expression and resulting genetic changes which can initiate cancer [Macfarlane and Murphy, 2010].

This gene regulation by miRNAs plays an important role for many major cell functions like growth, differentiation or metabolism and is currently examined by many investigators [Ardekani and Naeini, 2010]. The impor-

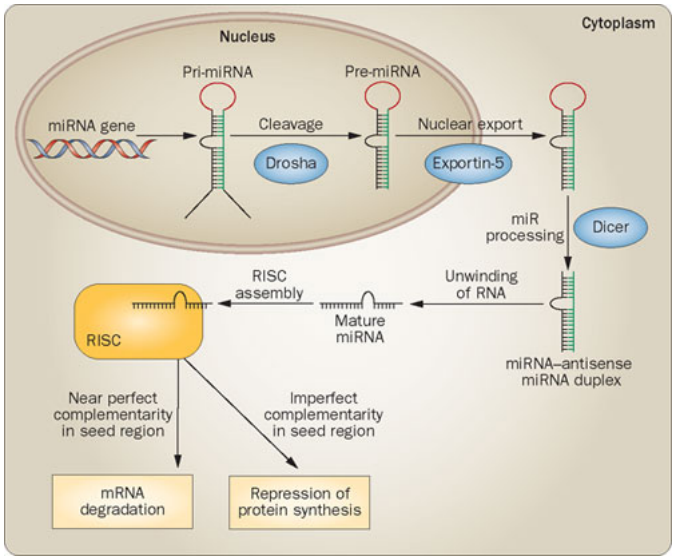


Figure 1.1: Biogenesis of miRNAs

tance of miRNA and their regulation can also be observed by the fact that more than 60% of all human genes have one conserved miRNA-binding site ore even more [Ha and Kim, 2014]. Therefore disregulation, either up- or downregulation, may result in different human diseases, e.g. cancer or autoimmune diseases [Ardekani and Naeini, 2010]. Hence, this also leads to the assumption that miRNAs are a useful feature for the diagnosis and treatment of diseases. But they follow a very special and complicated targeting and this has to be elucidated. So 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 to diagnose or treat diseases.

But as already mentioned, 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 [Enright et al., 2003]. 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 [Betel et al., 2010]. 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 [Lewis et al., 2005].

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 alignment 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 [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 are sequences of the precursor and of all mature miRNAs [Kozomara and Griffiths-Jones, 2011]. 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 [Karolchik et al., 2004].

2 Methods and Data

2.1 Methodology

only functional ones, not non and not weak input data and runtime, csv etc.

The program for the analysis is implemented in Python. As input files the UCSC gene file, miRTarBase file and miRBase file are required in csv format. Additionally the conversion file of gene id end nm number is needed. It can be given as a text file. The data files are parsed and each entry is stored into the respective dictionary that allows fast searching. Then for every MTI in the miRTarBase file the respective miRNA sequence is searched 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'. Here I only consider the strong validated data of the Functional MTIs because they deliver strong evidence for the interaction. So if an entry in the miRTarBase file contains something different than 'Functional MTI', it is scipped.

After matching these two sequences, a local alignment is performed with them. 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 [??, 2000a]. 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

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

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 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 [??, 2000b]. 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 [Enright et al., 2003]. The other parameters are just logically selected to test which influence they have on the results.

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. So for each found alignment list the miRNA name, the respective NM-number and the alignment score is 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 string is written to a text file.

For the analysis of the complementarities the alignment is further exam-

ined. Each position of miRNA and target mRNA is compared and whenever the nucleotide is similar, a '1' is added to a list, 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 for each position the number of '1's is computed and divided by the total number of elements in the matrix. So 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 and then aligned in the same way as the sequences of the true target interactions. This randomly assigment can by chance contain true targets but generally the miRNAs are matched with genes that are no true targets.

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 [??, 2000c]. 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.

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 postions were searched. For that only the MTI id was put into the url and then opened. Together with the corresponding miRNA name and target NM number the positions were stored in a text file. In the final table the found starting positions were compared to the positions given in the resulting text file above. The final number of found positions are stored in the final result table as well.

2.2 Prediction features

2.2.1 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 [Rehmsmeier et al., 2004]. The main complementary region or seed region includes nucleotides from position two to eight starting from the 5' end. Figure 2.1 shows the scheme of the seed region [Peterson et al., 2014]. This irregularity of the presence of non-complementarity regions makes the reliable prediction more difficult for animals than for plants. Because concentration 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 in the mRNA although there is no binding site at this position. It would be recommended to take more than the seed region into account to increase the complexity and specificity of the prediction. How significant the consideration of the sequence complementarity of miRNA and mRNA sequence is, will be further analysed with validated data in this research.

In the mentioned seed region there can be different types of matching patterns: perfect matching between six nucleotides, 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 [Lewis et al., 2005] [Brennecke et al., 2005] [Krek et al., 2005]. Figure 2.2 illustrates the different types of sites. Grimson et al. [Grimson et al., 2007] investigated the different types of sites referring to the effectiveness of the gene repression. As shown in figure 2.3, they found that the repression is the highest when an 8mer site is present and the lowest when only six nucleotides are perfectly matching.

2.2.2 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

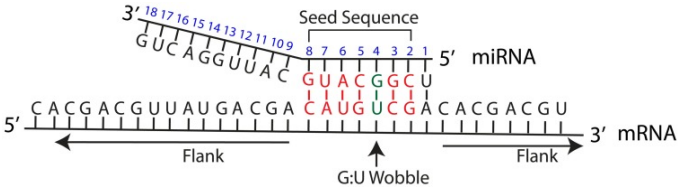


Figure 2.1: Scheme of seed matching

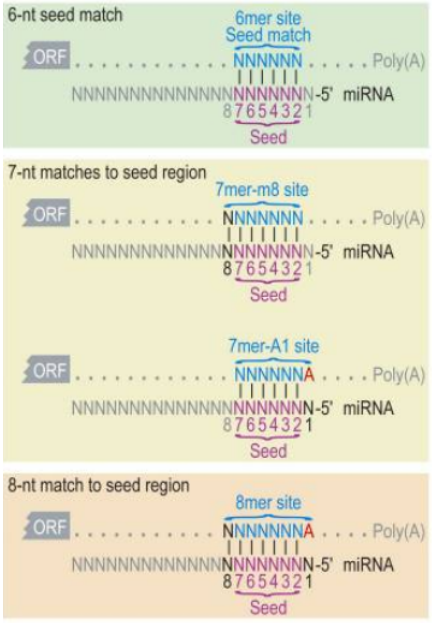


Figure 2.2: Canonical sites of seed region

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 [Rehmsmeier et al., 2004] rely mainly on this feature. (thermodynamics von miranda)

2.2.3 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 func-

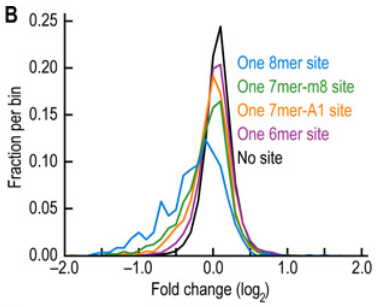


Figure 2.3: Effectiveness of different types of sites

tion [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. In terms of conservation, figure 2.6 shows how conserved the different nucleotides of a miRNA are implicating regions like the seed region and additional 3' pairing.

2.2.4 Site accessibility

Kertesz et al. [Kertesz et al., 2007] investigated the importance of site accessibility for the target prediction. 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. [Kertesz et al., 2007] found that if the targets form highly 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.

2.2.5 Addtional Watson-Crick pairing

In addition to the seed matching towards the 5' end, another complementary site towards the 3' end in the miRNA is present (Figure 2.4). Grimson et al. [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 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 2.6). 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 (Figure 2.7).

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].

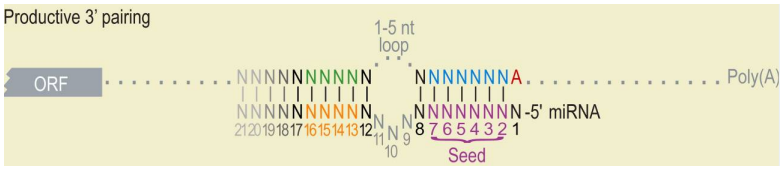


Figure 2.4: Additional 3' pairing

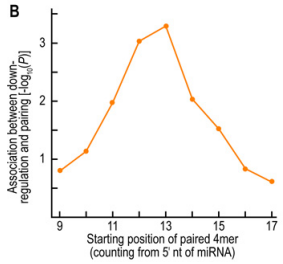


Figure 2.5: Relation between regulation and starting position of additional pairing

2.2.6 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

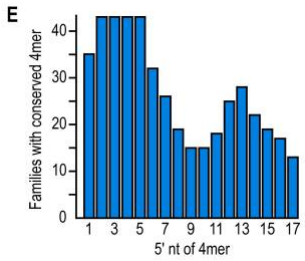


Figure 2.6: Conservation of nucleotide positions of miRNA

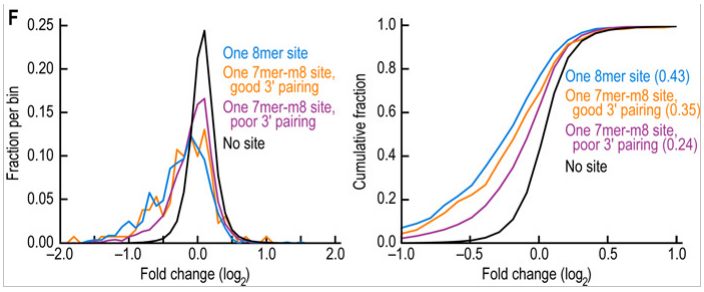


Figure 2.7: Efficacy of different combinations of seed region and additional 3' pairing

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 [Doench and Sharp, 2004]. 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 [Enright et al., 2003].

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

Enright et al. [Enright et al., 2003] and Doench et al. [Doench and Sharp, 2004] also found that the presence of multiple miRNA target sites results in a higher repression and destabilization of the mRNA. Grimson et al. [Grimson et al., 2007]

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 2.9). 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 2.8).

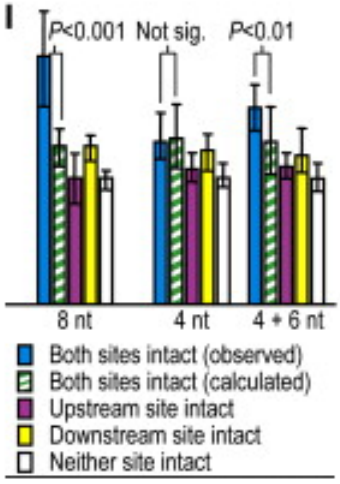


Figure 2.8: Cooperative repression with different site spacings

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 [Grimson et al., 2007].

Figure 2.10 shows the different efficacies. Another characteristic concerning the site locations is the distance from the stop codon. Figure 2.11

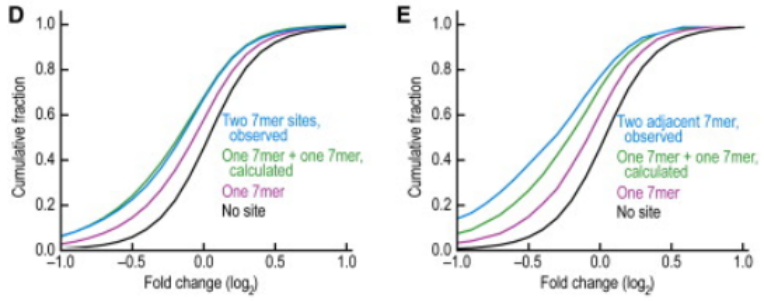


Figure 2.9: Effect of multiple sites

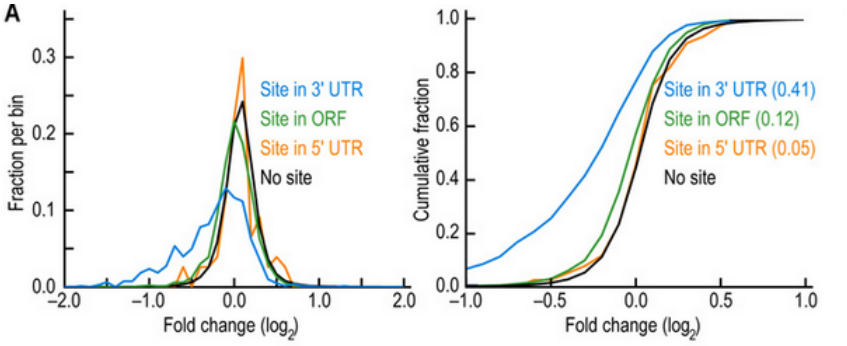


Figure 2.10: Efficacy of different site locations

illustrates that approximately in the first 15 nucleotides 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 [Grimson et al., 2007].

Considering all these different common and less common features, new targets for miRNAs can be roughly reliably predicted. Figure 2.12 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.

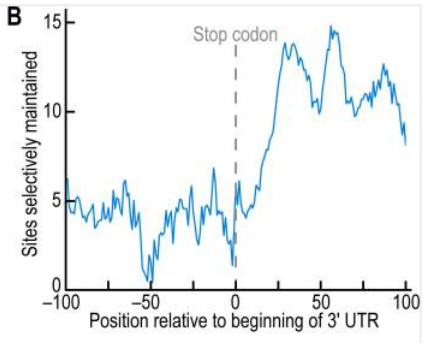


Figure 2.11: Efficacy of sites located relative to the stop codon

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 2.12: Prediction tools and their features

2.3 Data

2.3.1 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]. 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 [Hsu et al., 2011]. 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 above, 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.

2.3.2 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 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 [Kozomara and Griffiths-Jones, 2011].

2.3.3 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 [Karolchik et al., 2004]. 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 [Mudunuri et al., 2009]. 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 3.1 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 only 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 very 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 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,

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 scores of the true targets seem to be distributed in the higher scores and the ones of the non targets in the lower scores.

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 (No. 1 and no. 8) such a distribution was plotted which is shown in figure 3.1.

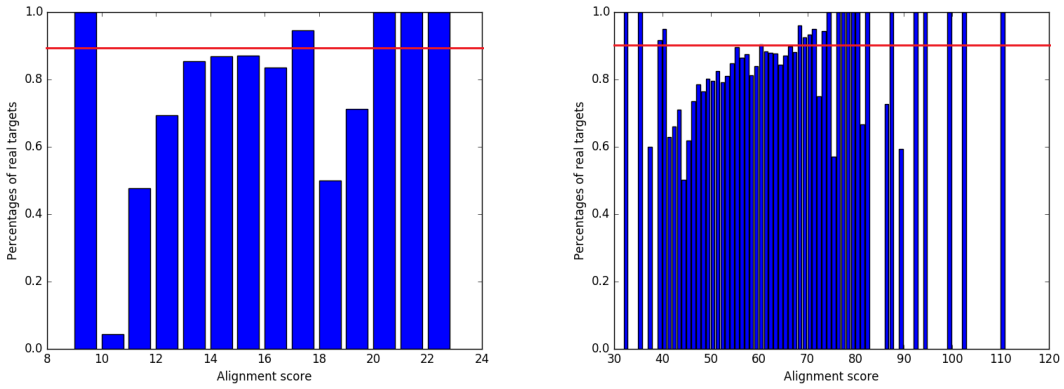


Figure 3.1: Distribution of alignment scores, parameters -4 -4 and 5 -3 -8- 2

On the x-axis the different existing scores of the respective data set 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. But also, there are some scores that are exceptions because more non targets produced this scores than true targets shown by lower y-values. Using this score as a prediction feature, a certain threshold needs to be determined to identify a true target. In figure 3.1 a threshold of 90% is delineated meaning at least 90% true targets produced a certain alignment but still 10% of random non targets had this score as well. Assuming this threshold and using the alignment scores for a prediction, for the first parameter set if we have scores of nine, 17, 20, 21, 22 or 23, a mRNA would be classified as a true target. Obviously the score nine would not fit in the pattern because it is a really low alignment score and the binding site would not be very suitable. On the other hand, alignments of miRNA and mRNA with a resulting score or 19 which belongs to the higher scores for this set, would not be classified as a true target even though the binding might be good. How this classification will look like depends strongly on this threshold. A lower threshold will result in a higher lower specificity and higher sensitivity, a higher one vice versa, so the two measurements of the performance need to be balanced.

For the second plot the same can be observed: some lower alignment score are above the threshold although they are not suitable and for some higher ones the probability to be a true target would be low.

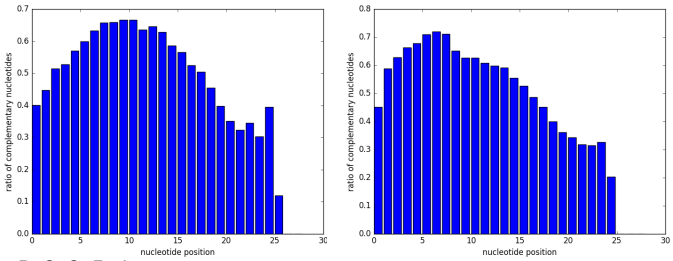
Summing up, assuming the score as the only prediction feature and given a certain threshold, some mRNAs are predicted right to be true targets but there are also false positives. Given high alignment scores, a true target is not necessarily present. This uncertainty can be compensated by considering many more features to combine them and get a more reliable prediction.

For each parameter set the complementarity of single nucleotides of the alignments of miRNA and target gene sequence were also plotted. These plots are shown in figure 3.2.

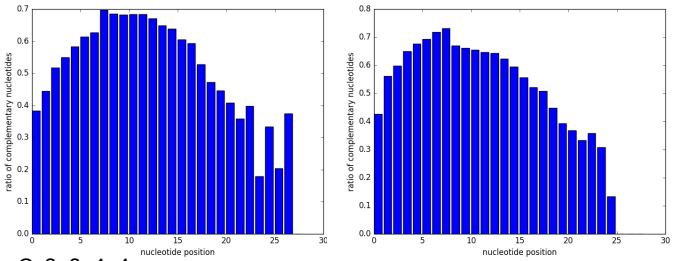
The x-axis of the plots shows the positions one to 22 or more of the miRNA, the y-axis the ratio of complementary bases at this position considering all alignments produced with these parameters. Because most of the miRNAs are only 22 nucleotides long, I will only focus on these positions because higher positions are very imprecise because of the different lengths.

For some sets there is not a significant difference between true targets and non targets, e.g. in plots H and I. Regarding set A, for the true target right an increase in complementarity in the seed region can be observed. On

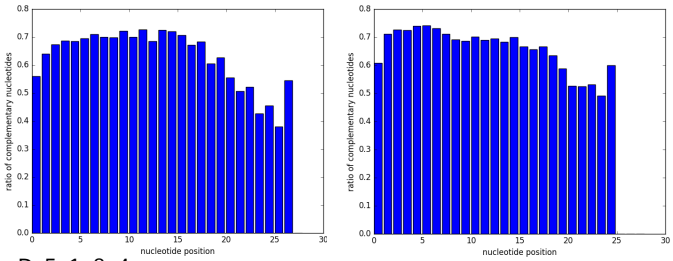
A: 1 -2 -2 -1



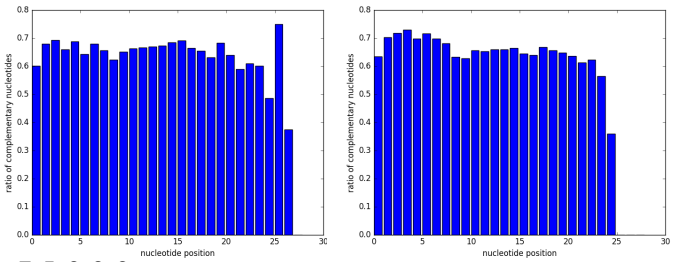
B: 2 -2 -5 -4



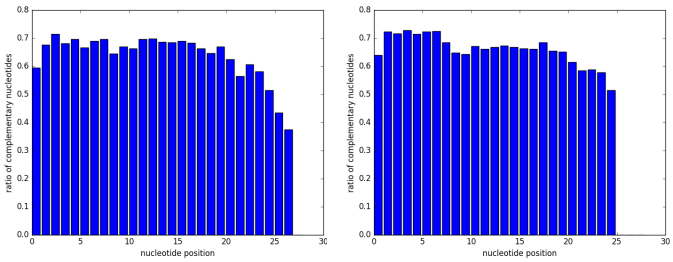
C: 3 -2 -4 -4



D: 5 -1 -8 -4



E: 5 -2 -8 -2



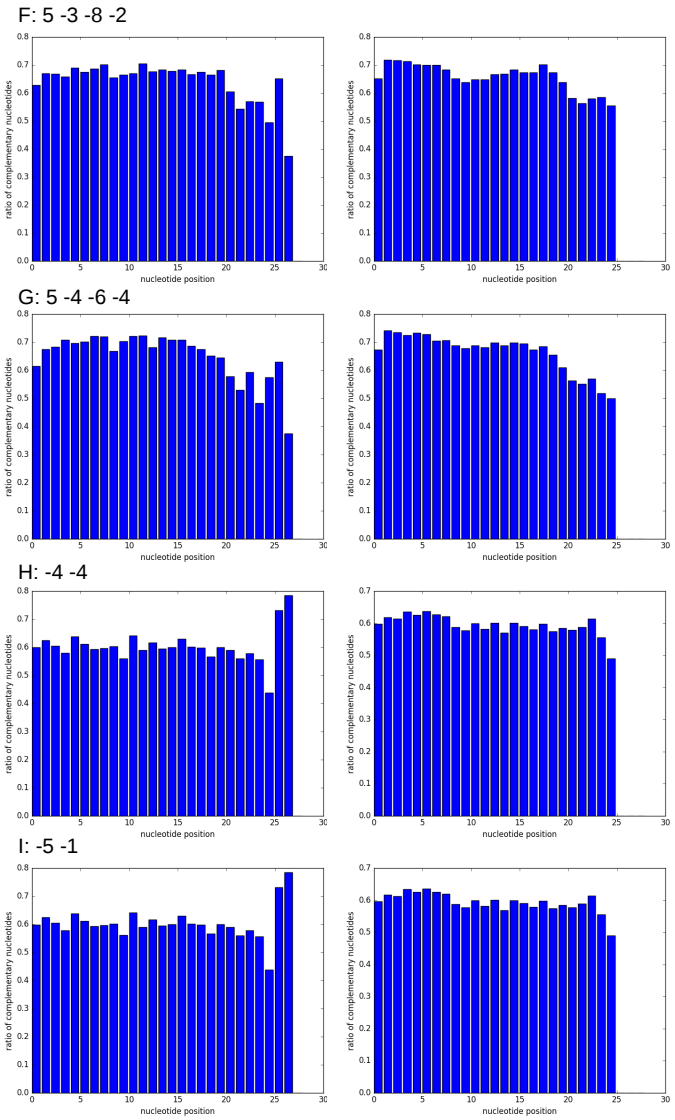


Figure 3.2: Ratios of complementarity per parameter set (A-H): targets right, non targets left

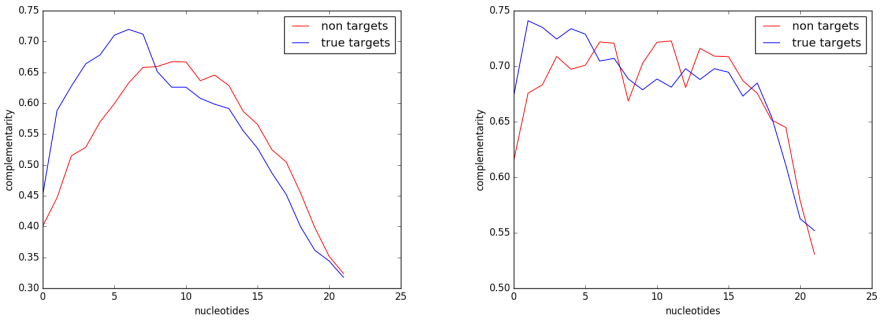


Figure 3.3: Curves of complementarity of non and true targets: 1 -2 -2 -1, 5 -4 -6 -4

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, the maximum of complementarity of the non targets left 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. The same applies for plot B. Considering the other figures of the complementarities there are no significant differences between the two groups. For the non targets all other plots, especially F to I show no big variability overall. The ratios are pretty similar comparing the positions. For the true targets a slight increase in the first positions can be seen but the increase amounts only to a few percentages. Also 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.

Other plots:

The analysis of the alignment starting positions was not really 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 (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 provided positions in the miRTarBase the numbers of the respective own found ones are in general not very high. This can be due

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

to the different sizes of the 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 resulting in different defined positions, although the alignment would start at the same position. 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.

Another reason for the inconsistent positions may be, that the tools they used in miRTarBase use better and more specific alignment tools where more parameters can be defined. The pairwise2 module only has 4 parameters and only delivers the alignment with the highest score and no ranking. In the database always the best three alignment positions are given.

The table in figure 3.4 shows that the average of the complementarity ratios are pretty similar comparing true targets with non-targets. Highlighted in yellow are the nucleotides of the seed region, in blue the nucleotides of the additional 3' pairing. Building the average ratios of these regions a possible increase could be observed. For the first parameter set, the true targets show a slight increase in the seed region but a decrease in the additional pairing. So the described feature from above can not be significantly observed. For the same set and the negative control, a constant average can be seen. For almost all true target an increase in the seed region and a decrease in the other region can be observed. For the non targets there are different patterns. E.g. for the third parameter set both regions are increased in contrast to the standard average. For the other sets either no big difference can be seen or both are enriched but never decreased. Going back to the plots, there the seed region in the true targets is slightly more visible than for the non targets. But all in all the differences in here are not very big and significant.

Both regions can not be observed as clearly as expected.

4 Discussion

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.

biomarker: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3199035/

As described above there are many features that can be considered and many different existing tools. 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 strength 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. [Peterson et al., 2014] the following tools are outstanding because of maintenance, newest input data and they are the easiest once to use: DIANA-microT-CDS, miRanda-mirSVR and Targetscan. All three are somehow unique but all e.g. use looser thresholds for conservation allowing less conserved regions to balance irregularities and refuse less true targets.

The early miRanda tools is also still widely used. The class of tools that use machine-learning are getting more accurate the more positive and negative target data is verified and features are found. Because of a lack of this data these tools are not significantly better than the three other tools mentioned above [Peterson et al., 2014].

The future consists of the elimination of the limits the tools and finding more useful features for the prediction.

Fujiwara and Yada [Fujiwara and Yada, 2013] followed a new approach by considering other characteristics than binding sites for the prediction, in fact 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 theirs method is almost as good as the standard binding site base ones, but combining the two different methods decreases the accuracy much. The advantages of this novel approach are independence of conservation of binding sites and the amount of available training data.

Coronnello and Benos [Coronnello and Benos, 2013] investigated another approach trying to improve the prediction power. They developed a tool ComiR that considers additionally 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. ?? www.ncbi.nlm.nih.gov/pmc/articles/PMC3692082/

n future die limits die noch da sind zu berwinden um besser zu werden. New tools: http://www.ncbi.nlm.nih.gov/pubmed/23445489/ combination of tools to overcome limits: http://www.ncbi.nlm.nih.gov/pubmed/23716633/ integrate expression data: http://www.ncbi.nlm.nih.gov/pubmed/23703208/

newst one:http://www.ncbi.nlm.nih.gov/pubmed/27561079 http://www.ncbi.nlm.nih.gov/pubm

http://www.ncbi.nlm.nih.gov/pubmed/27494513

http://www.ncbi.nlm.nih.gov/pubmed/27477696

http://www.ncbi.nlm.nih.gov/pubmed/27450903

http://www.ncbi.nlm.nih.gov/pubmed/27438777

References

[??, 2000a] ?? (2000a). Definition of sequence alignment. URL `https://en.wikipedia.org/wiki/Sequence_alignment`.

[??, 2000b] ?? (2000b). Documentation of module pairwise2. URL `http://biopython.org/DIST/docs/api/Bio.pairwise2-module.html`.

[??, 2000c] ?? (2000c). Documentation of ttest function of libreoffice. URL `https://wiki.openoffice.org/wiki/Documentation/How_Tos/Calc:_TTEST_function`.

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

[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.

[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.

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

[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.

[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/`.

[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/`.

[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.

[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/`.

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

[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.

[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`.

[O'Kelly et al., 2012] O'Kelly, F., Marignol, L., Meunier, A., Lynch, T. H., Perry, A. S., and Hollywood, D. (2012). MicroRNAs as putative mediators of treatment response in prostate cancer. *Nat Rev Urol*, 9(7):397–407.

[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.

List of Figures

List of Tables

Figure 1.1: [O'Kelly et al., 2012] http://www.nature.com/nrurol/journal/v9/n7/images/nrurol.
f1.jpg Figure 2.1: http://journal.frontiersin.org/article/10.3389/fgene.2014.00023/full
Figure 2.2: http://www.targetscan.org/docs/7mer.html
Figure sites: http://www.ncbi.nlm.nih.gov/pubmed/17612493
Table 2.12: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3927079/

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									
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 3.4: Table of ratios of complementarities