Agnes Leitner

MicroRNA target prediction

Bachelor Thesis

Institute for Genomics and Bioinformatics
Graz University of Technology
Petersgasse 14, 8010 Graz, Austria
Head: Univ.-Prof. Dipl.-Ing. Dr.techn. Zlatko Trajanoski

Supervisor:

Dipl.-Ing. Dr. techn. Hubert Hackl

Graz, July 2009

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…….. ……..
date (signature)

1 Contents

2 Abstract

MicroRNAs (miRNAs) are small noncoding RNAs that are involved in the regulation of protein expression in plants and animals. They are about 22 nucleotides long and they predominantly bind to the 3’ untranslated region of messenger RNAs to inhibit translation or to induce cleavage. Since the specific function of most miRNAs is unknown, it is necessary to find their target mRNAs. Because experimental identification of miRNA targets is difficult, several computational tools have been developed for predicting miRNA targets. Here the principles of target prediction and some prediction programs are presented.

3 Background

3.1 Introduction

miRNAs are small non-coding RNAs that are involved in the regulation of protein expression in plants and animals. They are about 22 nucleotides long and they predominantly bind to the 3' untranslated region (3'UTR) of messenger RNAs (mRNAs) to inhibit translation or to induce cleavage. It is thought that miRNAs can have hundreds of targets. The first miRNA lin-4 was discovered 1993 [1]. Until now over 9000 miRNAs in 103 species are known [2]. Most miRNAs in plants show near perfect complementarity to their targets. This feature facilitates the identification of miRNA-target interactions [3]. For miRNAs in animals the target prediction is more complex because very few miRNAs are perfectly complementary to their targets. In the following only animal miRNAs are considered.

3.2 Biogenesis

In most cases the genes of miRNAs do not lie near their target gene in the genome [4]. Consequently the transcription mechanism is completely independent and these miRNAs presumably have their own promoters. Additionally some miRNA genes are within a genomic cluster. Clustered miRNAs can be related to each other, but they do not have to share functional relationships. Accordingly, it is supposable that the clustered miRNAs are transcribed from one single primary transcript. Furthermore, there are also some miRNAs, which are in the introns of protein-coding host genes. These miRNAs are not transcribed separately but processed from the introns. Thus they have the same regulatory elements and primary transcript as their host genes [4].

The miRNAs in the introns are obviously transcribed by polymerase Pol II, as mRNAs are transcribed by Pol II as well. There are observations suggesting that many of the other miRNAs are also produced by Pol II, but there might still be some which are transcribed by Pol III [4].

Little is known about the transcription process but the primary miRNAs (pri-miRNAs) are thought to be much longer than the miRNA precursor, also known as pre-miRNA. The pre-miRNAs come from cleavage of the pri-miRNA by Drosha RNase III. The

pre-miRNA is ~60-70 nucleotides long and forms an imperfect stem loop structure. It is then transported from the nucleus to the cytoplasm. After that the enzyme Dicer cuts of the terminal base pairs and loop. The result is a miRNA:miRNA* duplex. It is composed by the mature miRNA and its complementary strand (miRNA*). But this duplex tends to be short-lived. When the miRNA is incorporated into the RNA induced silencing complex (RISC) the duplex is separated and the miRNA* is degraded [4].

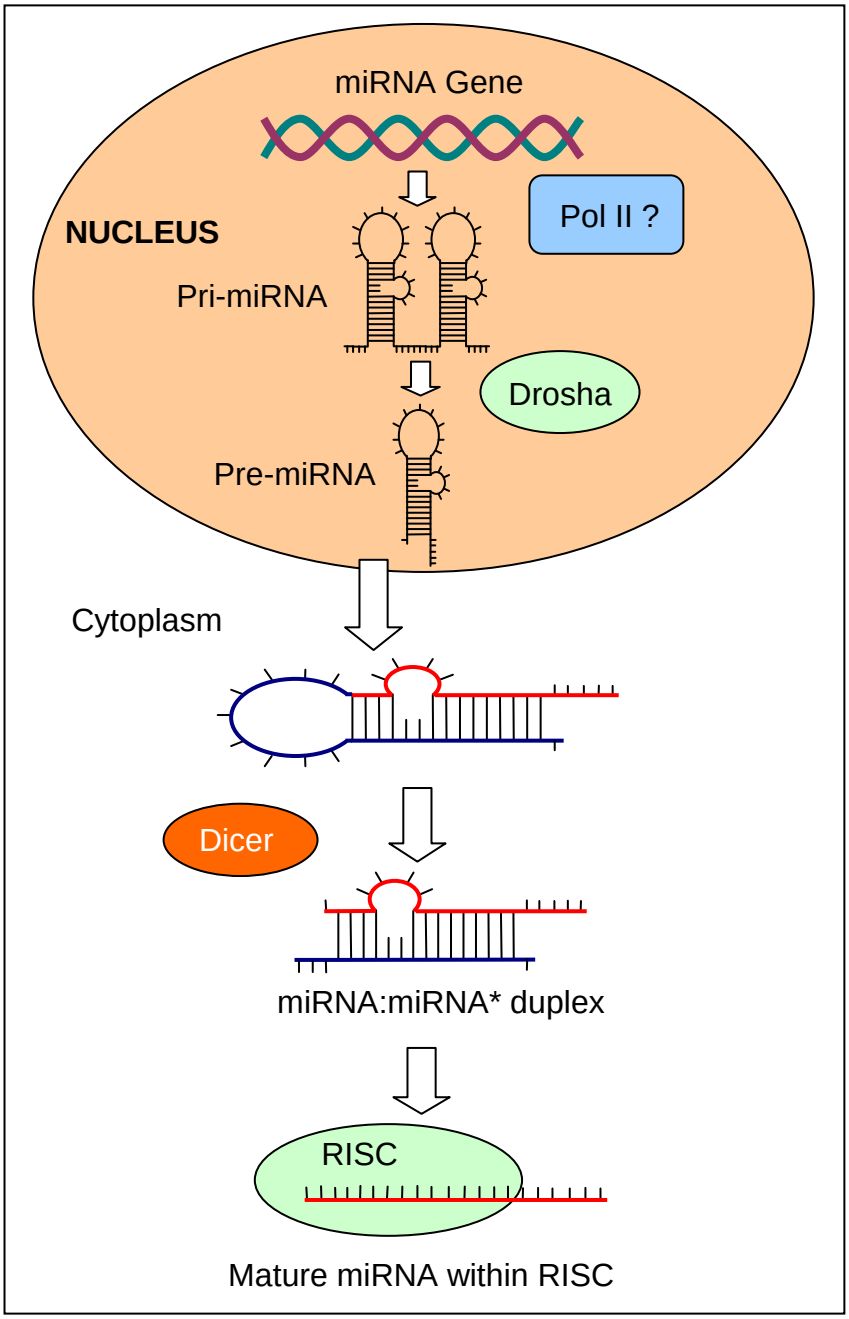

Figure 1: Biogenesis of metazoan miRNA (adopted from [5])

3.3 Function

miRNAs direct gene expression by two mechanisms. miRNAs bind to the RISC and guide it to cause either degradation of mRNAs or translational repression [6]. Whether the mRNA is cleaved or whether productive translation is inhibited depends on the complementarity of the miRNA to the mRNA. If there is a high degree of complementarity the RISC will cleave the mRNA. If the complementarity is not enough for cleavage but still fitting, translation will be repressed.

The RISC contains at least one Argonaute protein (Ago) that associates with the miRNA. The Argonaute family has several different members. Whether the miRNAs guide mRNA cleavage or whether they repress translation, also depends on into which specific Ago protein the miRNA is incorporated. Ago2 is thought to be responsible for RISC cleavage activity [7]. Nevertheless miRNAs are predominantly thought to act as translational repressors.

Furthermore an mRNA can contain multiple sites for the same or different miRNAs. Consequently several different miRNAs can act together to repress the same gene. It seems that these multiple target sites work independently. The response to multiple miRNAs increases nearly the same as if the responses to the single miRNAs for their own were multiplied [9, 17, 8].

miRNAs predominantly bind to sites in the 3'untranslated region (3'UTR) of their target mRNA. Nevertheless targeting can also occur in 5'UTRs and open reading frames (ORFs) [6, 9]. Although a significant amount of target sites has been found in ORFs, they seem to be less effective and are still less frequent than 3'UTR target sites. 5'UTR targeting is even rarer. This can be explained by the mRNA-clearing activity of the translation machinery. Since the ribosomes move from the cap-binding complex at the 5'side through the ORF to the 3'side of the mRNA, it is likely that silencing complexes bound to 5'UTRs or ORFs are rather displaced by the ribosomes than the ones bound to 3'UTRs.

4 Principles of miRNA target prediction

4.1 Sequence complementarity

At the 5'end of the miRNA there is a region called "seed". It is centered on nucleotides 2-7. Watson-Crick paring of the mRNA target site to this seed region seems to be the most important factor for miRNA target prediction [6]. Strictly requiring seed pairing improves highly the performance of miRNA target prediction tools. It reduces notably their false-positive rate. Additionally it is advantageous to favor A across nucleotide 1 (1A-anchor) of the miRNA over a Watson-Crick match. This is supported by analyses, which showed that this preference is conserved in vertebrates. Furthermore there is experimental evidence that these sites outperform others with a Watson-Crick match to position 1 [8, 10]. Most miRNA targets have a 7nt match. So either nucleotides 2-8 build Watson-Crick pairing (7mer-m8) or nucleotides 2-7 build base pairs combined with an A across position 1 (7mer-A1). Requiring perfect 8nt pairing (8mer) increases specificity, whereas searching for 6nt seed pairing (6mer) yields greater sensitivity.

Apart from seed pairing, pairing to the 3' end of miRNAs also plays a role in target recognition [6]. It can supplement seed pairing and consequently improves binding specificity and affinity. Such 3'-supplementary pairing ideally centers on miRNA nucleotides 13-16 and has a length of 3-4 nucleotides. However, these so called "3'-supplementary sites" are very rare and only have a modest effect.

Furthermore pairing to the 3' region of the miRNA can also compensate for a mismatch in the seed region [6]. These so called "3'-compensatory sites" are centered on miRNA nucleotides 13-17. The pairing can extend to nine consecutive Watson-Crick matches. However, 3'-compensatory sites are rare and presumably emerge only when a specific member of a miRNA family is required for regulation. Because miRNAs within a family have the same seed region but differ in their remaining sequence.

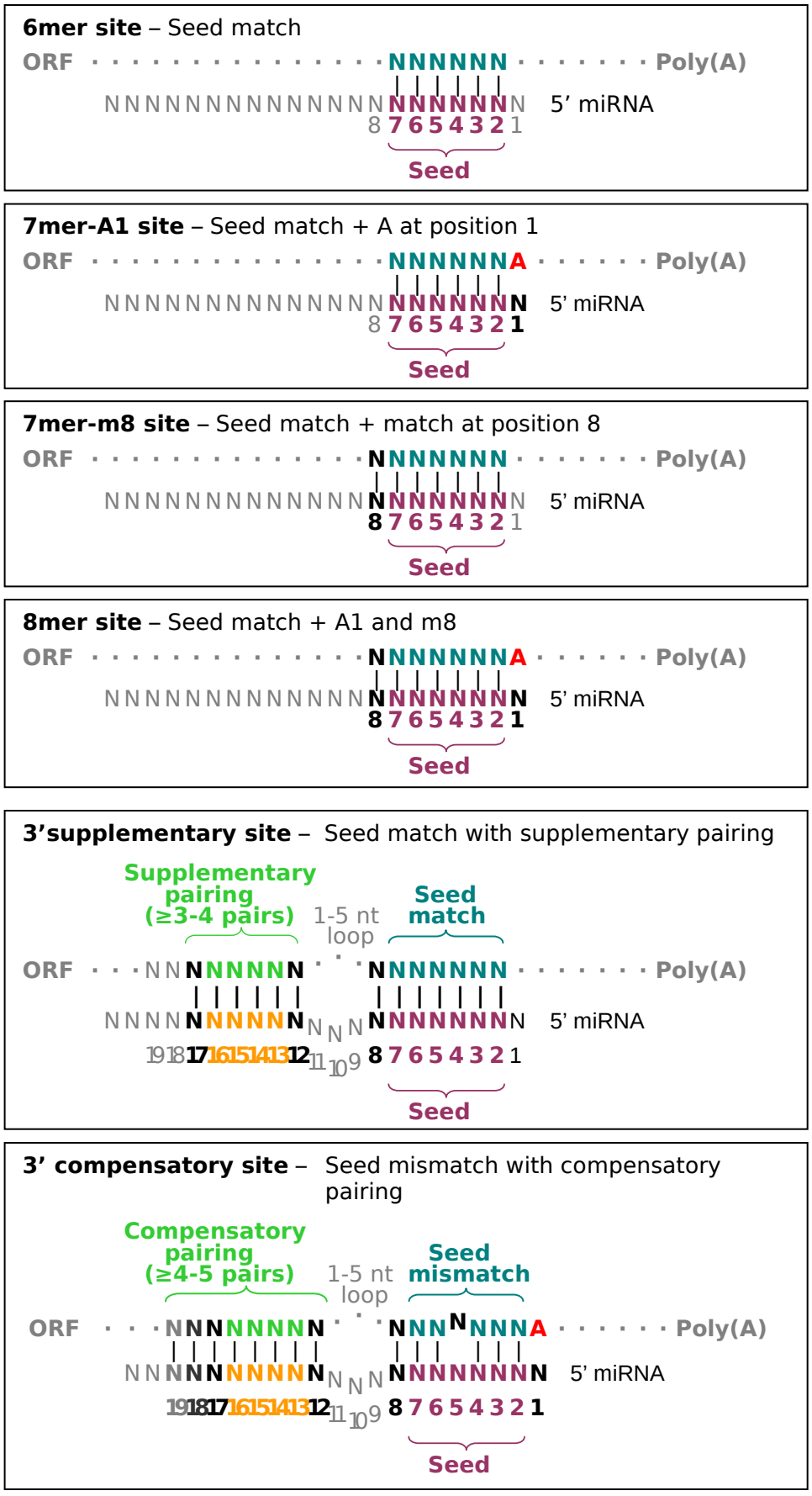

Figure 2: Types of miRNA target sites (adopted from [6])

But why is the seed region of miRNAs so important? This can be explained by the way how the miRNA is bound by the silencing complex. For efficient pairing it seems to be ideal when the RISC presents nucleotides 2-8 of the miRNA preorganized in the shape of an A-form helix to the mRNA. Other configurations appear to result in lower affinity and efficiency [4, 11].

4.2 Conservation

Since binding sites that are conserved across species are likely to be biologically functional, these sites are potential miRNA target sites. The use of conserved site sequences reduces the false-positive rate of prediction programs significantly.
However, the diverse prediction programs sometimes use slightly different definitions of conservation [12]. Sites are commonly regarded as conserved if they are retained at orthologous locations in multiple genomes, which means they have to appear exactly at the same position in the alignment of the 3'UTR sequences. Sometimes it is sufficient when the regions corresponding to the seed of a miRNA fall in overlapping alignment positions [13]. Nevertheless sites can also be regarded as conserved if they just can be found somewhere in the sequences but not in aligned positions [14]. Additionally it is possible, that the site is missing or has changed in only one of the multiple organisms that are considered. These sites can be regarded as poorly conserved.

4.3 Thermodynamics

Another approach for identifying miRNA targets is the consideration of thermodynamic stability. A commonly used parameter is the free energy of the miRNA:target duplex, ΔG_{duplex}.
It is an energetically more favorable state when two complementary RNA strands are hybridized. The lower the free energy of two paired RNA strands, the more energy is needed to disrupt this duplex formation. Thus, a RNA duplex is in a thermodynamically more stable state, which means the binding of the miRNA to the mRNA is stronger, when the free energy is low (more negative). Consequently a miRNA has a higher affinity to bind to an mRNA, when the following duplex has a low free energy ΔG_{duplex}.

4.4 Site Accessibility

The secondary structure of the mRNA appears also to play an important role. For binding to the miRNA the target site has to be accessible, which means it has to be opened and must not interact with other sites within the mRNA [15, 16]. This opening brings an energetic cost ΔG_{open}, which also has to be considered. Additionally it seems favorable when short regions with a length of ~15 nucleotides upstream and downstream of the target site are opened as well. The total free energy change $\Delta\Delta G$ equals to the difference between ΔG_{duplex} and ΔG_{open} and represents a score for the accessibility of the target site and the probability for a miRNA-target interaction.

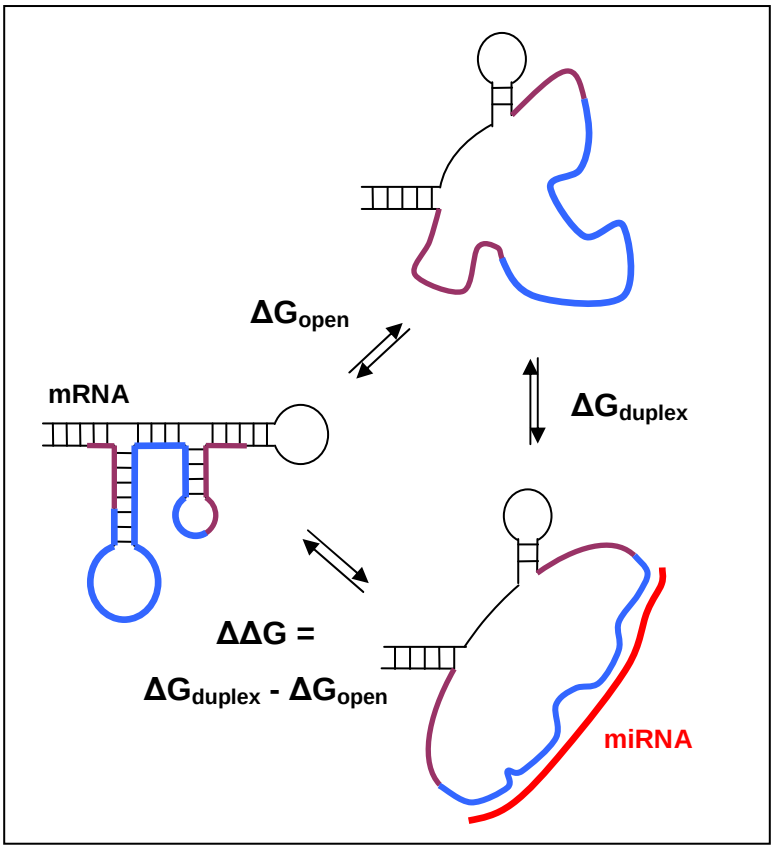

Figure 3: Illustration of free energy changes for miRNA-target interactions

4.5 UTR Context

It is thought that not only the sequence of the target site defines whether an mRNA is target of a certain miRNA, but also the UTR context [6, 17]. For instance the position of the site influences efficacy. Accordingly, the site has to be located within the 3'UTR at least 15nt from the stop codon. In long UTRs it should not fall in the middle,

because at this location the site might be less accessible to the silencing complex. Furthermore high local AU content seems to increase accessibility of a site because of the weaker mRNA secondary structure. These assumptions are supported by conservation analysis of 7-mers in orthologous 3'-UTRs [17, 18].
Additionally proximity to binding sites of coexpressed miRNAs enhances site efficacy. For two sites that are close together act cooperatively. Which means it leads to a greater response than expected by the normal multiplicative effect of multiple sites, when two sites are within 40nt but not closer than 8nt [17, 19].

4.6 Correlation of expression profiles

It has been shown that by transfecting miRNAs into cells plentiful mRNAs are downregulated, which indicates that these mRNAs are targets to the individual miRNAs [22]. Hence, genes that are lowly expressed within a tissue in which a specific miRNA is naturally expressed are likely targets to the miRNA. By this inverse relationship between the expression profiles of miRNA and their predicted targets it is possible to draw conclusions about the interaction of a miRNA and a potential target. To identify miRNA-target interactions that result in mRNA degradation, an analysis of mRNA and miRNA levels is sufficient. To determine targets of miRNAs that act as translational repressors, data about protein levels is required.

4.7 Validation of prediction tools

The different miRNA target prediction programs yield quite different results due to the different approaches that were used and the varying implementations. Subsequently it is difficult to decide which predicted miRNA-target interactions are more likely to be accurate and which programs provide the best performance, respectively.
There are some possibilities to estimate the correctness of prediction tools.
Because miRNA-target interactions are either functional or non-functional it is reasonable to determine sensitivity (eq.1) and specificity (eq.2).

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positves} + \text{false negatives}} \qquad (1)$$

Sometimes sensitivity is also called true positive rate (TPR). "True positives" (TP) is the number of predicted miRNA-target interactions that really exist. "False negatives"

(FN) is the number of miRNA-target interactions that do exist, but were not predicted. Thus, sensitivity is the relation of the number of predicted interactions to the number of all existing interactions.

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{false postives}} \tag{2}$$

“True negatives” (TN) is the number of non-existing miRNA-target interactions that correctly were not part of the predictions. “False positives” (FP) is the number of interactions that were erroneously predicted, but actually do not exist. Subsequently, specificity is the relation of the number of correctly not predicted interactions that do not exist to the number of all non-existing interactions.
Often instead of specificity the false-positive-rate (FPR) is determined (eq.3).

$$\text{False positve rate} = \frac{\text{False postives}}{\text{False positves} + \text{true negatives}} = 1 - \text{Specificity} \tag{3}$$

To achieve a good performance of a prediction tool, both sensitivity and specificity have to be maximized. Which means the predictions should contain not much false positives and as few as possible false negatives. Defining more tolerant thresholds for example for free energies or conservation scores, results in higher sensitivity but also in a loss of specificity. Hence, it is necessary to find the best trade-off of these two measures.

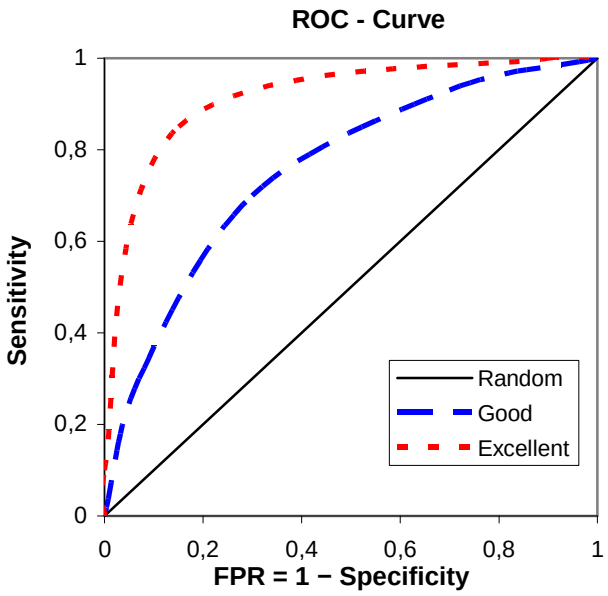

Figure 4: ROC – Curve

A suitable method to optimize the relation of sensitivity and specificity is a ROC (receiver operating characteristic) analysis. The ROC curve is a plot of the sensitivity versus the false-positive rate or (1 − specificity) (Figure 4). It helps to find the optimal balance between sensitivity and specificity. Furthermore, the area under the curve (AUC) is a comparable value for the over-all performance. The bigger the AUC is, the better the performance of the method. If the AUC equals to 0.5 the results are worthless (like random).

However, to determine sensitivity and specificity it is necessary to have a data set with a sufficient number of experimentally verified and refuted miRNA-target interactions. These interactions should be unbiased. Which means they should not be discovered by means of particular prediction programs [20].

Another approach to evaluate the significance of the results of a prediction tool is to estimate the signal-to-noise ratio (SNR) of the target predictions. This is done for example by the use of “mock” miRNAs [21]. “Mock” miRNAs are random sequences that are individually designed for each existing miRNA. They have approximately the same number of seed matches in 3’UTRs as their real correspondent [12]. These mock sequences are unlikely to be biologically relevant, consequently prediction programs usually predict fewer conserved target sites for the mock miRNAs than for real miRNAs. This indicates that many predicted conserved target sites for real miRNAs are indeed biologically functional.

In short, the ratio of the number of predictions for real miRNAs (“signal”) to the number of predictions for mock miRNAs (“noise”) is an estimate of the SNR. However, a low SNR does not imply that the predicted target sites are false, they just cannot be distinguished from noise [12].

4.8 Experimental Verification

Although the current prediction tools help finding miRNA targets, they still lack sensitivity and specificity [20]. Therefore it is necessary to verify the predictions experimentally.

The most common method for testing miRNA-mRNA interactions is the reporter gene assay, which provides direct evidence about the functionality of a miRNA:mRNA pair. The expression of a reporter gene can be easily identified. The Green Fluorescent Protein (GFP) and Luciferase are often used as reporter genes. GFP fluoresces, when expressed in a cell and exposed to blue light, and Luciferase causes

bioluminescence. Therefore it is easy to observe the expression of these genes. For the purpose of examining miRNA-target interactions, the 3'UTR of the observed mRNA is attached downstream of the reporter gene. Then it is introduced into a cell line of interest. By measuring the expression level of the reporter gene in absence and presences of a specific miRNA, it is possible to draw conclusions about the fact whether the miRNA is targeting the 3'UTR.

To determine whether the miRNA induces translational repression or mRNA cleavage, it is necessary to measure mRNA levels, as well. This is most commonly done by means of reverse transcription polymerase chain reaction (RT-PCR).

Further techniques to observe miRNA:mRNA interactions are microarray analysis [22] and pSILAC (pulsed stable isotope labeling with amino acids in cell culture) [9]. However, these methods only provide indirect evidence, because they just detect changes in expression profiles and not the direct interaction of an individual miRNA to a specific mRNA. Microarrays measure changes of mRNA levels. Thus they detect just miRNA:mRNA interactions that cause mRNA cleavage and degradation. pSILAC is a method that directly measures changes in protein production induced by overexpression of miRNAs [9]. Therefore it also recognizes downregulation of gene expression by translational repression.

Another approach to verify a miRNA-target interaction would be to knockout the miRNA gene and examine the effects on protein changes [6, 8]. However, because of the high number of targets for a miRNA a miRNA-target interaction cannot be assigned with confidence by this way. Therefore more sophisticated methods to disrupt only one specific miRNA-target interaction are required [6].

Since the argonaute protein makes close contacts to miRNAs and their targets, high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) can be used to identify miRNA-mRNA interactions [23]. By immunoprecipitation of Ago it is possible to gain the miRNAs and the mRNA binding sites to which Ago forms crosslinks.

5 miRNA target prediction tools

Several tools have been developed to predict targets of miRNAs. There are differences in the used approaches and implementations. In the following some of these tools will be described. Table 1 gives an overview of current miRNA target prediction tools.

5.1 miRanda

The target predictions of microRNA.org are based on the miRanda algorithm [24, 25]. miRanda analyzes the complementarity between a given mRNA and a set of miRNAs using a weighted dynamic programming algorithm. It computes a weighted sum of scores for matches and mismatches of base pairs. G:U wobbles are allowed but less scored than perfect matching base pairs. The weights are positiondependent. Since complementarity in the seed region seems to be the most important factor for miRNA targeting, scores for base-pairing at positions 2-8 have a greater weight. Additionally base-pairs in the 3' region are also weighted higher to regard for example 3'-compensatory matches. Further the free energy of the miRNA:mRNA duplex is estimated by using the Vienna RNA folding package [26] and used as filter. Moreover conservation of the miRNA:mRNA relationship is considered to filter out less conserved predicted targets. For that purpose the PhastCons conservation score [27] is used.

At microRNA.org it is possible to work online and to download the data as well. The website provides precompiled predictions of miRNA targets for human, mouse and rat. It is possible to search online either for the targets of a given miRNA or for the miRNAs that are targeting a specific gene. The results are then displayed in the web browser. Additionally sets of predictions are provided for download.

Furthermore the miRanda source code is freely available. The miRanda algorithm scans one or more given miRNA sequences against a set of given RNA- or DNA sequences to find potential target sites.

5.2 miRNA - target prediction at EMBL

The method of EMBL searched for miRNA targets in flies [28]. The 3'UTRs were scanned for 8nt to 4nt complementarity to the seed region of the miRNA. For 8mers one mismatch or loop and for 7mers one G:U wobble were allowed. The matches had to be 100% conserved within two species of drosophila to be further considered. Then an individual score for each remaining site was determined. Therefore the pairing energies of the miRNA 5'end and 3'end to the mRNA were calculated separately using RNAhybrid [29]. If the seed-matches contained mismatches or if they contained less then 7 consecutive base-pairs then an appropriate 3' pairing energy was required to compensate this. For instance, for 7mers with a G:U wobble 60% of the maximally possible paring energy at the 3'end was needed [28].
To obtain more comparable scores than the free energies, Z scores were calculated for the 5' and 3' pairing energies (Z = standard deviations above the mean of background matches) [30]. Subsequently these scores were added, whereas the 5' scores were weighted accordingly to the seed-match type [28]. The resultant site scores were summed to gain the UTR score.
The results of this method for fly are provided at www.russell.embl-heidelberg.de/miRNAs/.
For each used miRNA all predicted targets are listed in one file that is available for download. Additionally it is possible to download the individual target sites.

5.3 PicTar

The method PicTar („probabilistic identification of combinations of target sites") does not only identify putative targets for single miRNAs but also ranks target genes by considering whether the mRNA is target for combinations of miRNAs [13]. In each cell type different miRNAs are coexpressed. This suggests that these sets of miRNAs regulate tissue-specific target genes. Hence, input to PicTar is such a set of miRNAs and a group of orthologous 3'UTRs from multiple species. PicTar then determines common targets for the miRNAs and ranks them by their likelihood of being a target.
To detect targets for single miRNAs PicTar requires perfect 7mer seed matches (perfect Watson-Crick pairing of either nucleotide 1-7 or 2-8). Additionally imperfect seed pairings are allowed but they typically do not contribute much to the PicTar score. The results are then filtered by evaluating the free energy of the miRNA:mRNA

formation using RNAhybrid [29]. Further the sites are checked for conservation to reduce the false positive rate. Then to each remaining site a probability is assigned. It corresponds to their likelihood of being functional [13].
These values are input to the PicTar Sequence Scoring Algorithm, which computes a maximum-likelihood score using a Hidden Markov Model (HMM). The scores are calculated for each species individually. The results are combined to get a final PicTar score, which describes the likelihood of a gene being target to the given miRNA set [13, 31].
At http://pictar.mdc-berlin.de/ predictions for mice, vertebrates, flies and worms are available. The results are provided by an online search interface. By entering a microRNA ID or a Gene ID it is possible to search for the miRNAs' targets and for all miRNAs that are predicted to target the gene, respectively.

5.4 TargetScan and TargetScanS

TargetScan.org provides different approaches for predicting microRNA target sites in several species.
The first version of TargetScan basically searched for seed pairing and ranked the resulting sites by evaluating thermodynamic stability. The results for multiple species were combined to get the predictions for conserved target sites [21]. Later a simplified version called TargetScanS was published [32]. This method just looked for pairing to a 6-nt miRNA seed with an additional base pair at nucleotide 8 or a 1A-anchor. In addition it was also required that the seed matches were conserved at conserved positions. Other criteria as thermodynamic stability were not included to this approach. Recently an improved version of TargetScan was published. In particular a new method for evaluating site conservation, which performed considerably better, was introduced [33]. Target sites with imperfect seed matches but 3' compensatory pairing are now also predicted. In mammals the efficiencies of the sites are evaluated just by observing the UTR context of the target sites [17].
www.targetscan.org provides miRNA predictions for human, mouse, rat, dog, chicken, chimpanzee, rhesus, cow, opossum, frog, worm and fly. The online search interface provides the opportunity to ask for targets of a miRNA by specifying the miRNA name or by selecting the miRNAs' family from a given list. It is optional whether conservation is considered or not. Additionally a search for the targets of a given gene is implemented.

Furthermore it is possible to download the target predictions. A Perl script for identifying miRNA targets is also available.

5.5 DIANA-microT 3.0

The algorithm of DIANA-microT searches in the UTRs for stringent seed paring (at least 7 consecutive Watson-Crick pairs) to the miRNA. 6mers or seed matches containing one G:U wobble are also accepted as putative target sites when they are compensated by pairing to the 3'end of the miRNA. Subsequently the conservation of the sites and the binding type are considered for scoring each site. In addition the putative target sites are compared with sites identified based on mock sequences to gain scores that include miRNA-specific SNR and the estimation of a precision score. The total score of a target is then calculated by building a weighted sum of the individual scores of each target site on the 3'-UTR [34].
Predictions of DIANA-micro-T for human and mouse are available at http://diana.cslab.ece.ntua.gr/microT/. The website provides the opportunity to search for both, targets of a specific miRNA and miRNAs targeting a defined mRNA. DIANA-microT additionally provides a signal-to-noise ratio and a precision score for each result, to give the possibility to assess the "correctness" of the predictions.

5.6 PITA

PITA is the short for "Probability of Interaction by Target Accessibility". As the name indicates, this tool primarily identifies miRNA targets by considering the accessibility of target sites within the mRNA [15]. It first scans the 3'UTRs for putative target sites. This is done by searching for near-perfect seed matches. For each site ΔΔG is computed. It is optional whether the flanking sequences upstream and downstream of the target site are considered for determining ΔΔG. It appears to be ideal when 3 nucleotides upstream and 15 nucleotides downstream are included for the calculation. For computing ΔΔG existing programs that predict secondary structures were used. ΔΔG represents a score for the probability of a miRNA-target site interaction. When there are multiple sites for one miRNA the ΔΔG scores are appropriately summed to gain a score for the total interaction energy for the miRNA:target pair. Thus, PITA also includes the number of target sites for scoring and ranking miRNA targets.

On the one hand PITA provides a better sensitivity and specificity than other algorithms, although it does not consider for example conservation. It appears to be more accurate and simplified than other existing methods [15]. On the other hand it is questionable whether the programs for secondary structure predictions are suitable and whether the prediction of pure RNA structure is useful, as the real mRNA structure can differ from predictions because of RNA-binding proteins and several other aspects [16, 6].

The PITA web site (http://genie.weizmann.ac.il/pubs/mir07) provides files containing miRNA target predictions for mouse, fly, worm and human for download. It is further possible to search for miRNA targets online. The online search interface provides the opportunity to define the required seed type and a conservation threshold and to select whether the sequences flanking a potential target site should be considered. Running the PITA algorithm on a specific UTR sequence or miRNA sequence is either possible by using the online prediction tool or by downloading and executing the executables of PITA, which are available at the web site.

5.7 ElMMo

Generally this method searches for seed pairing and then evaluates the functionality of the sites by conservation analysis based on a Bayesian probabilistic model [18].

First putative target sites for each miRNA are determined by scanning the 3'UTR for complementarity to 3 different seed types. More precisely it is looked for 8mers, and the two types of 7mers, whereas at the mRNA position across nucleotide 1 of the miRNA an A is not accepted as a match, instead Watson-Crick pairing at position 1 is required. However, the miRNAs 3'end is not observed, consequently 3' compensatory paring does not contribute to this method. To each potential target site a posterior probability, that the site is target of the miRNA, is assigned. Therefore the method determines a conservation pattern for each site separately within a set of related species and models the evolution of the site [18].

At http://www.mirz.unibas.ch/ElMMo3/ the latest miRNA target predictions of ElMMo released 2009 are available. Predictions for worm, zebrafish, fly, human, mouse and rat are provided. By selecting miRNAs from a given list it is possible to query the target prediction for them. Additionally text files containing all predictions for one organism are available for download.

5.8 mirWIP

mirWIP is a method that considers the free energy of the miRNA:mRNA duplex, seed pairing, conservation and site accessibility [14].

mirWIP uses RNAhybrid [29] with modifications to identify possible miRNA:target matches in 3'UTRs. This is done in a very liberal manner. Consequently the predictions are quite imprecise and need to be filtered. While running RNAhybrid the results are already filtered by the free energy of the duplex. miRNA:target duplexes with a free energy value above a certain limit are discarded. Next the pairing to the miRNA 5'end is observed. mirWIP does not require stringent seed paring. Consequently 6mers containing G:U wobbles and bulges in 7mers are accepted. Further perfect base-pairing has not to be centered on nucleotides 2-7 but is also allowed to begin at position 3. In addition matches at nucleotide 1 or A-anchors are not regarded.

The following conservation filter uses a relaxed definition of conservation. It requires for the individual putative target sites just a match in the orthologous UTRs, but the matches do not have to lie within aligned blocks. The remaining sites are used as the initial set for further observations.

The predictions of the initial set are then scored by evaluating three features: 5' seed matching (S), structural accessibility (A) and total energy (E). To each of these characteristics a scoring parameter is assigned. These parameters are derived from the analysis of target predictions based on immunoprecipitation of the RISC components AIN-1 and AIN-2 [14]. The total site score is the product of these values:

$Score_{site} = S \times A \times E$

The accessibility (A) of the remaining sites is analyzed using Sfold [35]. Sfold computes for each nucleotide in the 3'UTR the probability, that it remains unpaired. Based on this probabilities an average accessibility of the site and surrounding sequences is calculated.

The total interaction energy (E) is calculated in a separate step from the average accessibility. The total energy $\Delta\Delta G$ equals to $\Delta G_{duplex} - \Delta G_{open}$.

In a final step scores for the individual miRNA families are determined by adding up the scores of all nonoverlapping sites of each member. If these values are below 2 the binding sites of the entire family are discarded. To obtain a total target score the contribution from each remaining miRNA family is summed.

The results for the target prediction in C.elegans are available at mirtargets.org. The search interface provides the opportunity to search either by miRNAs or by mRNAs for the predictions. The source code for mirWIP is also online available as supplementary software for [14]. Further, mirWIP has been integrated into the STarMir module (http://sfold.wadsworth.org/starmir.pl). It provides the possibility to make predictions for any miRNA-target pair online.

5.9 rna22

rna22 is a pattern-based method for miRNA target prediction. It makes use of miRNA patterns to identify “target islands”, regions of the mRNA sequence that have a higher likelihood to contain binding sites for miRNAs [36].

First statistically significant patterns were derived from a set of known mature miRNAs. The patterns were generated using the Teiresias algorithm [37]. Next it is searched for sequences in the 3’UTR that correspond to these patterns. When many miRNA patterns cluster around a specific UTR location this region is called “target island” and it is associated with a putative miRNA binding site. In the next step the miRNAs that will bind to these “target islands” have to be identified. Therefore all putative target sites are paired with all miRNAs and the structures and free energies of these miRNA:target duplexes are predicted using the Vienna package [38]. To restrict the number of the results the user can define three parameters. First the maximal free energy (E), the minimum number of base-pairs between the miRNA and the target (M) and the maximum number of unpaired bases of the miRNA seed region (G), whereas G:U wobbles count as matches. Note that rna22 does not consider conservation.

Although rna22 uses a limited set of miRNAs to train the algorithm, the method is able to predict target sites for miRNAs, which were not part of this set.

The implementation of rna22 is available at http://cbcsrv.watson.ibm.com/rna22.html. There are also precompiled sets of predictions for human, mouse, fly and worm provided.

5.10 GenMiR++

GenMiR++ (Generative model for miRNA regulation) is a tool that combines results of other miRNA target prediction programs with paired miRNA-mRNA expression profiling [39].
GenMiR++ uses target predictions of TargetScanS and associates them with miRNA and mRNA expression profiles from the same sets of tissues and cell types. Then it scores each miRNA-mRNA pair by a Bayesian approach. It evaluates whether the expression of the miRNA explains the expression levels of the mRNA. The profiles of all other miRNAs predicted to target the same mRNA are also considered. Target pairs achieve a higher score when the miRNA is highly expressed in tissues, in which the mRNAs were downregulated, and are further rewarded when other miRNAs, which potentially regulate the mRNA, are less expressed. In contrast the score of a target pair is reduced when both the miRNA and the mRNA are highly expressed in the same tissue [39]. Note, GenMir++ only predicts miRNA-mRNA pairs that result in transcript degradation. To consider interactions that lead to translational repression protein expression data would be needed.
At http://www.psi.toronto.edu/genmir/ it is possible to download MATLAB code and data for running GenMiR++. The list of human miRNA-target interactions scored by GenMir++ is also available for download (Supplementary Table 2 for [39]).

5.11 TarBase

TarBase is a database that provides information, whether a miRNA-target interaction is functional or not [40]. It contains about 1300 entries describing whether a miRNA-target interaction was tested positive or negative. 1000 entries concern human genes. This information was extracted from about 200 different scientific papers. Positive tested interactions are not only described by the miRNA and the targeted mRNA, but also by the type of the experiment used to validate the interaction and whether the miRNA causes translational repression or cleavage.
TarBase is accessible at http://diana.cslab.ece.ntua.gr/tarbase/.

5.12 miRBase

miRBase comprises three services concerning miRNAs. Up to date the database miRBase:Sequences provides information about over 9000 miRNAs from 103 species, including the sequences and genomic locations of the miRNAs [2]. The data is available through a searchable web interface or in form of downloadable files of various formats. miRBase:Targets provides miRNA target predictions. It uses the miRanda algorithm [24] with varied parameters to identify putative target sites. At the website a online search for target genes and miRNAs is provided as well as precompiled lists of predictions for download. The predictions are done for fly, worm and several vertebrates. miRBase:Registry is a service that assigns names to novel miRNA genes according to a defined miRNA nomenclature.
The services of miRBase are available at http://microrna.sanger.ac.uk.

5.13 miRGen - Targets

miRGen is a database that provides several services concerning miRNAs, including the opportunity to query predictions of multiple target predictions tools via a single interface. The database comprises predictions of DIANA-microT, miRanda (microrna.org), miRanda (miRBase), TargetScan and PicTar for the organisms human, mouse, rat, worm, fly and zebrafish. Additionally the experimentally supported targets from TarBase are included in the database. Moreover it is possible to query the union and intersections of the predictions of some programs [41].
miRGen is available at http://www.diana.pcbi.upenn.edu/miRGen/. On the one hand the website provides the opportunity to search for the targets of particular miRNAs or for the miRNAs targeting a specific mRNA. All common target gene ID types are supported by the online search. On the other hand files containing all predictions are available for download.

Tool	Method for Prediction and Ranking	Organism	Data availability	Executables availability	Website	Ref.
miRanda	Moderately stringent seed pairing, free energy, conservation	Human, mouse, rat	Online search, download	Yes	microRNA.org	[25]
EMBL	Seed paring, conservation, free energy	Fly	download	No	www.russell.embl-heidelberg.de/miRNAs	[28]
PicTar	Stringent seed pairing, free energy, conservation, probability being target to set of miRNAs	Mouse, vertebrates, fly, worm	Online search	No	pictar.mdc-berlin.de	[13], [31]
TargetScan	Stringent seed pairing, conservation, UTR context	Vertebrates	Online search, download	Yes	targetscan.org	[33], [17]
TargetScan	Stringent seed pairing, conservation	Fly, worm	Online search, download	Yes	targetscan.org	[32]
DIANA-microT 3.0	Seed pairing, conservation, site number	Human, mouse	Online search	No	diana.cslab.ece.ntua.gr/microT	[34]
PITA	Seed pairing, site accessibility, total interaction energy, site number	Human, Mouse, fly, worm	Online search, download	Yes	genie.weizmann.ac.il/pubs/mir07/	[15]
ElMMo	Stringent seed pairing, conservation	Human, mouse, fly, worm, zebrafish	Online search, download	No	www.mirz.unibas.ch/ElMMo3/	[18]
mirWIP	Moderately stringent seed pairing, free energy, conservation, site accessibility, total interaction energy, site number	Worm	Online search	Yes (Suppl. Software [14])	mirtargets.org	[14]
rna22	miRNA patterns, free energy, seed pairing	Human, fly mouse, worm	Online search, download	No	cbcsrv.watson.ibm.com/rna22	[36]
GenMir++	Relationship of miRNA and mRNA expression profiles	Human	Download	Yes	www.psi.toronto.edu/genmir	[39]
miRBase: Targets	Target predictions using miRanda algorithm with varied parameters	Vertebrates, fly, worm	Online search, download		microrna.sanger.ac.uk	[2]
Databases for miRNA genomics						
miRBase: Sequences	Database containing miRNA Sequences,				microrna.sanger.ac.uk /	[2]
TarBase	Database of experimentally verified miRNA-target interactions				diana.cslab.ece.ntua.gr/tarbase/	[40]
miRGen-Targets	Database comprising predictions and combined predictions of several tools				www.diana.pcbi.upenn.edu/miRGen/	[41]

Table 1: miRNA prediction tools

6 Discussion

6.1 Comparison of prediction tools

Many computational approaches for miRNA target prediction have been published. But which of them have the best performance and how reliable are their results? This question is difficult to answer. There are only few surveys that independently compare some of the tools.

A recent survey of Selbach and colleagues [9] analyzed 7 different tools. They compared the predicted targets of the programs TargetScanS, PicTar, rna22, PITA, miRBase, miRanda and DIANA-mircroT 3.0 with the results of a pSILAC analysis. More precisely they investigated the predicted mRNAs for five different miRNAs (miR-1, miR-16, miR-30a, miR-155, let7b) from each tool. Then they compared the number of mRNAs predicted by a specific tool which were measured with pSILAC as well with the fraction of these mRNAs that showed a $\log_2$-fold change lower than −0.1. The results are shown in Figure 5.

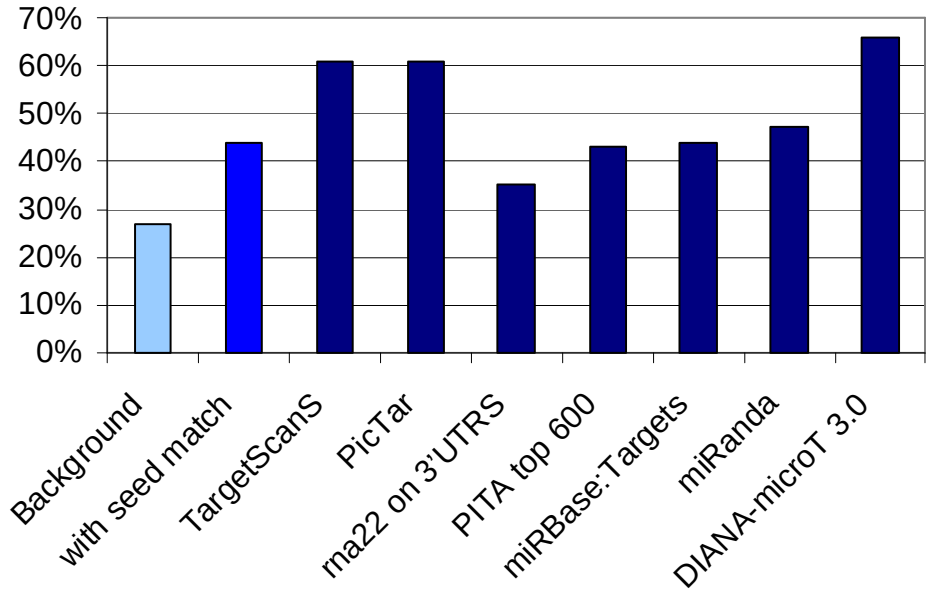

Figure 5: Fractions of predicted targets with reduced protein production (values taken from [9])

For the comparison of the results of PITA just the 600 top ranked predictions for each miRNA are considered. 27% of a completely random selection (Background) from the mRNAs considered in the pSILAC data show downregulation. This accuracy was

topped by all prediction programs. Considering only seed matches for predicting reveals a hit rate of 44%. This is clearly exceeded by TargetScanS, PicTar and DIANA-microT 3.0, which can be explained by a more rigid conservation filter [9].
These results are supported by an older survey by Sethupathy and colleagues [20]. They observed the mammalian target prediction programs DIANA-microT [42] (which is an older version of DIANA-microT 3.0), miRanda [43], TargetScan [21], TargetScanS [32] and PicTar [13]. Based on the experimental supported miRNA-target interactions provided by TarBase they estimated the performance of each program by determining sensitivity. They revealed a relatively low sensitivity for TargetScan and DIANA-microT. miRanda, TargetScanS and PicTar showed almost the same sensitivity of about 65%. However, miRanda predicted a considerably higher number of miRNA-targets interaction compared to the number of predictions of TargetScanS and PicTar, which indicates a lower sensitivity for miRanda. Sethupathy and colleagues further showed that the union of the results by some predictions programs yield higher sensitivity but also result in a higher number of total target predictions, whereas the intersection of the programs lead to a lower number of prediction but also drops the sensitivity. This indicates a great diversity of the predictions between the programs.

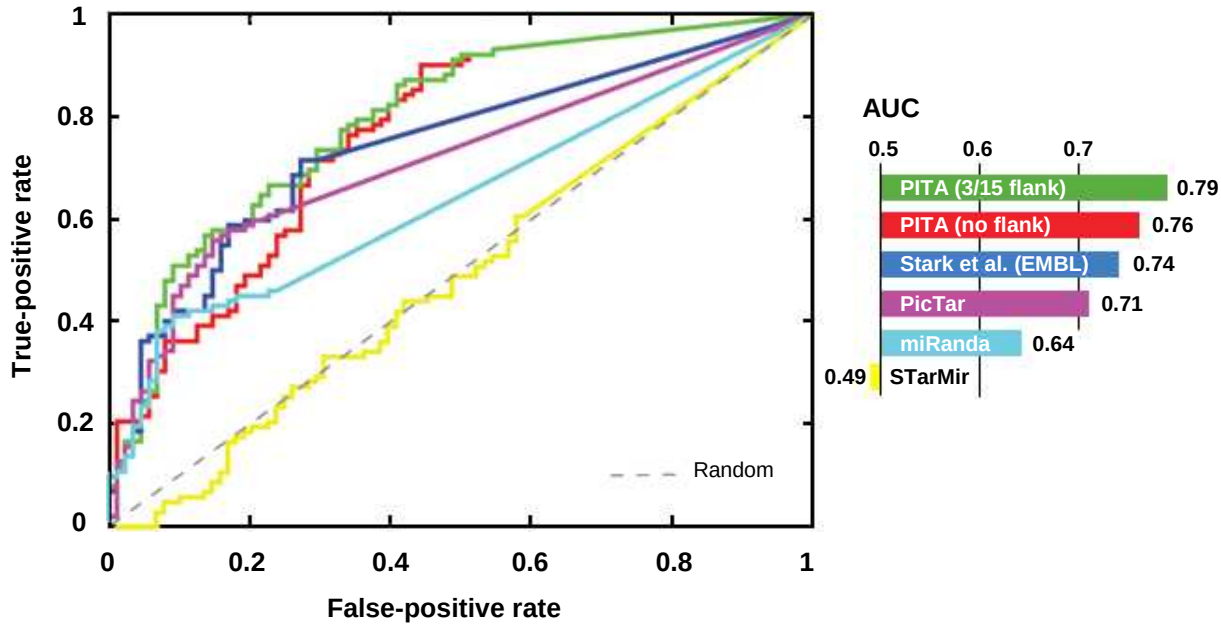

Figure 6: True-positive rate (sensitivity) versus false-positive rate (1 − specificity) for each score prediction threshold available for the specific prediction method. (The ROC-Curve and the AUC values were taken from [15].)

Another survey analyzed the results for fly of five prediction methods based on 190 experimentally tested miRNA-target interactions [15]. Sensitivity and specificity were calculated for PITA (considering flanks of 3 nucleotides upstream and 15 nucleotides downstream), PITA (considering no flanks) [15], the predictions of Stark et al. (EMBL) [28], PicTar [44], miRanda [43], and STarMir [45]. The predictions of each tool were sorted by score and subsequently sensitivity and specificity were plotted for each possible score threshold. The area under the curve (AUC) was also computed for each method (Figure 6). According to the AUC values PITA (with flanks) achieves the best performance followed by the method of EMBL and PicTar. However, the differences in performance are not very significant. Nevertheless it is notable, that PITA shows a comparable performance to the method of EMBL and PicTar, although it does not apply a conservation filter.

Baek and colleagues compared the results of seven prediction tools with the results of their examination [8]. They did a gene knockout concerning the miRNA mir-223 in mouse and monitored the changes of protein levels. Afterwards they tested the target predictions of miRBase:Targets [2], miRanda [24, 25], PicTar [13, 31], PITA [15] and TargetScan [17, 32] with their data. TargetScan and PicTar proved to be the most effective tools among those methods, which considered conservation. Moreover, the ranking of TargetScan correlated with protein downregulation by far most significantly [8]. This indicates that evaluating the UTR context of a potential binding site is very promising.

Tools that require stringent seed pairing perform better than those that do not [8, 9]. Additionally, the current predictions of these tools (TargetScan, PicTar, EMBL, ElMMo) have a high degree of overlap [6]. Baek and colleagues further showed that the tools, which do not consider conservation, do not perform better than a simple search for seed-matches. This result is also supported by the surveys described above [9].

A combination of the predictions by two or three different tools might be the most effective approach to get accurate results. For instance, the intersection of the predictions by PicTar and TargetScan would be reasonable, because they appear to be quite accurate prediction methods and they have a high overlap of about 80-90% [12]. Additionally considering the predictions of DIANA-microT would be probably a good approach, since it shows also a very good performance.

6.2 Example: Predictions for the gene MYCN

To illustrate the efficiency of just searching for seed matches a script in Perl was developed (see Appendix). The script detects perfect complementarity of the 5'end of a miRNA to a given 3'UTR sequence. Input is a list of miRNAs in FASTA-format and the 3'UTR sequence of one mRNA. The seed type is optional. The seed types 6mer, 7mer-A1, 7mer-m8 and 8mer are supported. To compare the results with the predictions of common prediction programs, first 8mer seed matches to the gene MYCN in human were computed (37 miRNAs were predicted). The gene MYCN was chosen because for this gene a verified miRNA-target interaction for the miRNA miR-101 was listed in TarBase. The 3'UTR sequence of MYCN was looked up in the UCSC Genome Browser. The human miRNA sequences were downloaded from miRBase [2]. Then the positions of the computed binding sites were plotted in addition to the positions of the miRNA binding sites predicted by some prediction tools (Figure 7a). For the plot the miRNAs that were predicted to target the human gene MYCN were investigated from the prediction methods PicTar [31], miRanda, TargetScan, DIANA-micro-T, PITA and ElMMo. The predictions of PITA were queried with the default seed type settings, a conservation threshold of 0.3 and under the consideration of the flanks of the sites. For the predictions of PicTar conservation among 5 mammals was required. From each prediction method the 35 – 38 best ranked miRNAs were selected, choosing a value about 36, because the most restrictive data set of PicTar provided only 36 predictions. The predicted binding sites for the miRNAs were marked. Further the verified binding sites of miR-101 that were obtained from TarBase were included. The binding sites of the miRNAs that were predicted to target MYCN by at least five of the six prediction tools are also marked (Intersection). The predicted binding sites of the miRNAs miR-101 and miR-98 were highlighted. Additionally the basewise conservation of the 3'UTR sequence and the alignment of the sequence in 44 vertebrates are shown (Figure 7b).

It is remarkable that in the region, which shows less conservation, clearly fewer target sites are predicted. The verified miRNA miR-101 was predicted by each method (green markers). The miRNA miR-98 was predicted by PicTar, TargetScan, DIANA-microT and ElMMo (red markers). Further there was a relatively low overlap between the predictions. 83 different miRNAs were predicted and top ranked by the diverse prediction tools to target the gene at about 50 different target sites. Only three,

including miR-101, of the 83 miRNAs were predicted by all prediction programs and only 12 miRNAs were predicted by at least 5 methods to bind at 6 different sites.

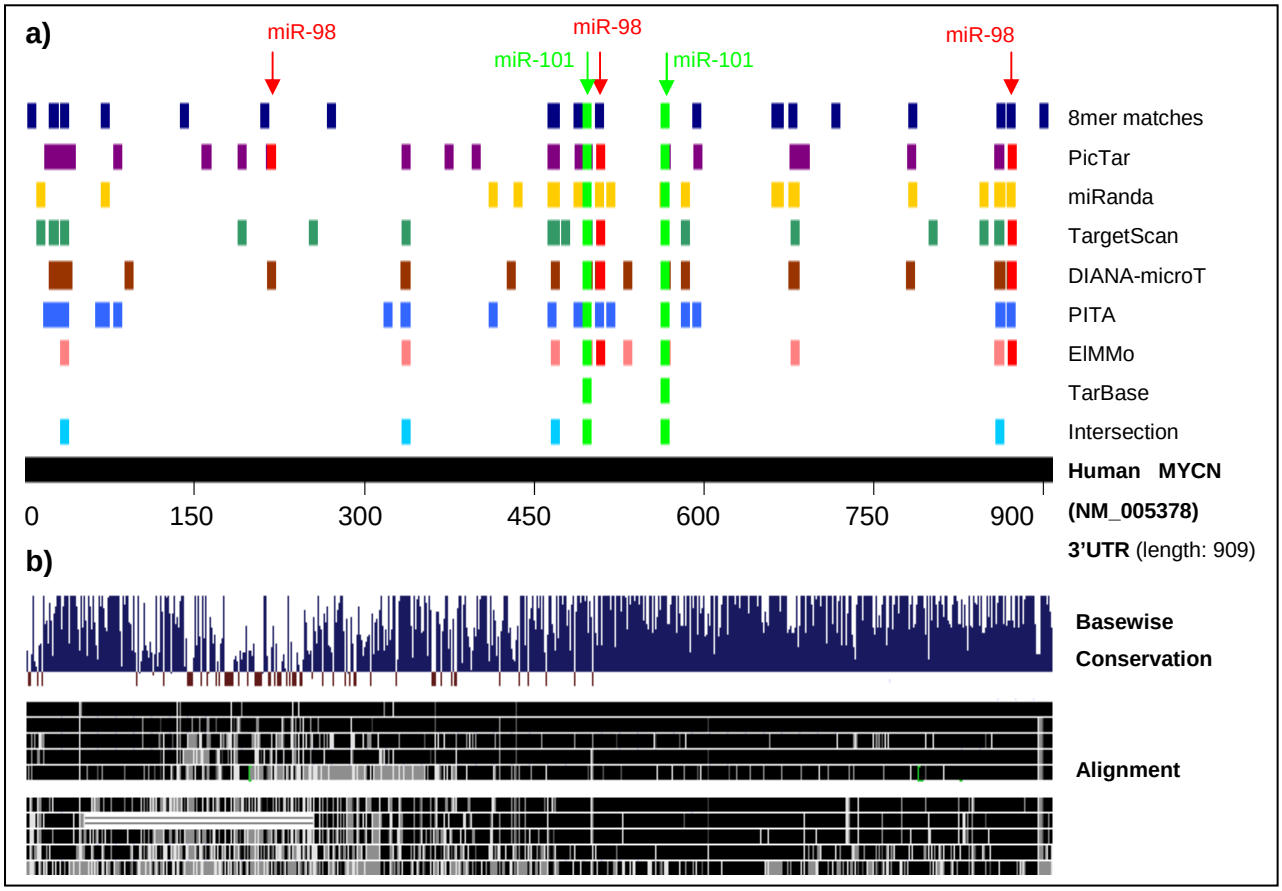

Figure 7: a) Plot of the binding sites on the gene MYCN of the seed region of the predicted and best ranked miRNAs by six different target prediction tools. Additional markers: verified binding site (TarBase), Intersection (miRNA binding sites that were predicted by at least five of the six prediction programs) and 8mer seed matches computed by the developed Perl script. Green markers: predicted binding sites of miR-101. Red markers: predicted binding sites of miR-98. **b)** Basewise conservation of the 3'UTR sequence and Alignment of 44 vertebrates (this graphic was taken from the UCSC Genome Browser)

Although the simple search for 8mer seed matches identified the binding sites of miR-101, the results of this method have a very low overlap with the results of all other prediction tools. More than 40% of all miRNAs that have 8mer seed matches are not predicted and top ranked by any other program, whereas not more than 25% of the results of any other prediction tool are single predictions. This low overlap of the 8mer matches is not surprising, due to the lack of any other filter like conservation of free energy.

7 References

[1] Lee RC, Feinbaum RL, Ambros V. The C elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 75: 843-854 (1993)

[2] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 36: D154-D158 (2008)

[3] Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP. Prediction of plant microRNA targets. Cell. 110: 513-520 (2002)

[4] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 116: 281-297 (2004)

[5] Brown JR, Sanseau P. A computational view of microRNAs and their targets. Drug Discov Today. 10: 595-601 (2005)

[6] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 136: 215-233 (2009)

[7] Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell. 15: 185-197 (2004)

[8] Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 455: 64-71 (2008)

[9] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature. 455: 58-63 (2008)

[10] Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA. 13: 1894-1910 (2007)

[11] Mallory AC, Reinhart BJ, Jones-Rhoades MW, Tang G, Zamore PD, Barton MK, Bartel DP. MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. EMBO J. 23: 3356-64 (2004)

[12] Rajewsky N. microRNA target predictions in animals. Nature Genetics. 38: S8-S13 (2006)

[13] Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, Macmenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N. Combinatorial microRNA target predictions. Nat Genet. 37: 495-500 (2005)

[14] Hammell M, Long D, Zhang L, Lee A, Carmack CS, Han M, Ding Y, Ambros V. mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nat Methods. 5: 813-819 (2008)

[15] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 39: 1278-1284 (2007)

[16] Hofacker IL. How microRNAs choose their targets. Nat Genet. 39: 1191-1192 (2007)

[17] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell. 27: 91-105 (2007)

[18] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics. 8: 69-69 (2007)

[19] Saetrom P, Heale BS, Snove O Jr, Aagaard L, Alluin J, Rossi JJ. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res. 35: 2333-2342 (2007)

[20] Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods. 3: 881-886 (2006)

[21] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 115: 787-798 (2003)

[22] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 433: 769-773 (2005)

[23] Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature. [Epub ahead of print] (2009)

[24] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA Targets. PLoS Biol. 2: e363-e363 (2004)

[25] Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNAorg resource: targets and expression. Nucleic Acids Res. 36: D149-D153 (2008)

[26] Wuchty S, Fontana W, Hofacker I, Schuster P. Complete Suboptimal Folding of RNA and the Stability of Secondary Structures. Biopolymers. 49: 145-165 (1999)

[27] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15: 1034-1050 (2005)

[28] Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell. 123: 1133-1146 (2005)

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

[30] Stark A, Brennecke J, Russell RB, Cohen SM. Identification of Drosophila MicroRNA Targets. PLoS Biol. 1: E60-E60 (2003)

[31] Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P, Kao HL, Gunsalus KC, Pachter L, Piano F, Rajewsky N. A Genome-Wide Map of Conserved MicroRNA Targets in C elegans. Curr Biol. 16: 460-471 (2006)

[32] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 120: 15-20 (2005)

[33] Friedman RC, Farh KK, Burge CB, Bartel D. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19: 1-11 (2009)

[34] Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P,

Hatzigeorgiou AG. DIANA-microT web server: elucidating microRNA functions through target prediction., Nucleic Acids Res. 37: W273-W276 (2009)

[35] Ding Y, Chan CY, and Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA. 11: 1157 – 1166 (2005)

[36] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell. 126: 1203-1217 (2006)

[37] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics. 14 55-67 (1998)

[38] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer SL, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures. Chemical Monthly. 125:167-188 (1994)

[39] Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD. Using expression profiling data to identify human microRNA targets. Nat Methods. 4: 1045–1049 (2007)

[40] Papadopoulos GL, Reczko M, Simossos VA, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res. 37: D155-D158 (2009)

[41] Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res. 35: D149-D155 (2007)

[42] Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. A combined computational-experimental approach predicts human microRNA targets. Genes Dev. 18: 1165-1178 (2004)

[43] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 5: R1-R1 (2003)

[44] Grun D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N. microRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Comput Biol. 1: e13-e13 (2005)

[45] Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. Nat Struct Mol Biol 14: 287-294 (2007)

8 Appendix

Perl script:

```
#Agnes Leitner
#MicroRNA target prediction
#Find seed matches in 3'UTR

#read input files and get seed-type

print "Enter name of FASTA-file containing miRNA sequences: ";
$input_mirnas = <STDIN>;
print "Enter name of FASTA-file containing 3'UTR sequence: ";
$input_mrna_sequence = <STDIN>;
print "Enter seed type (6mer/7mer-A1/7mer-m8/8mer): ";
$seed_type = <STDIN>;

until (($seed_type eq "6mer\n")||($seed_type eq "7mer-A1\n")||
          ($seed_type eq "7mer-m8\n")||($seed_type eq "8mer\n")||
          ($seed_type eq "quit\n")||($seed_type eq "q\n")||
          ($seed_type eq "exit\n"))
  {
    print "Seed type not supported!\n";
    print "Enter seed type (6mer/7mer-A1/7mer-m8/8mer): ";
    $seed_type = <STDIN>;
  }

$seed_type =~ s/[\n\r]//g;

$results = "seed_matches_$seed_type.txt";

#search for seed match

if(($seed_type eq "6mer")||($seed_type eq "7mer-A1")||
  ($seed_type eq "7mer-m8")||($seed_type eq "8mer"))
{
  searchForMatch($seed_type, $input_mirnas,
                 $input_mrna_sequence, $results);
  print "The results were written to the file $results.\n";
}
else
{
  print "Aborted!\n";
}

#-------------------------------------------------------------------------
#The sub "searchForMatch" searches for seed matches of a given seed type
#Input is the seed type, a list of miRNAs and the 3'UTR sequence

sub searchForMatch
{
  my (@parameter) = @_;

  my $seed_type = $parameter[0];

  #Define some parameters conserning the seed type accordingly to the
  #given seed type.
```

```
if ($seed_type eq "6mer")
{
  $seed_start = 1;
  $seed_length = 6;
  $A_anchor = 0;
}
elsif ($seed_type eq "7mer-A1")
{
  $seed_start = 1;
  $seed_length = 6;
  $A_anchor = 1;
}
elsif ($seed_type eq "7mer-m8")
{
  $seed_start = 1;
  $seed_length = 7;
  $A_anchor = 0;
}
elsif ($seed_type eq "8mer")
{
  $seed_start = 1;
  $seed_length = 7;
  $A_anchor = 1;
}
else
{
  print "seed type not supported!\n";
  return;
}

#open given files

open (MIRNAS, "$parameter[1]") or die
                              "I couldn't get at $parameter[1]";
open (RESULT, ">$parameter[3]")or die
                              "I couldn't overwrite $parameter[3]";
open (MRNA,$parameter[2]) or die "I couldn't get at $parameter[2]";

#extract sequence of mRNA

$seq_mrna = "";
for $line_mrna (<MRNA>)
{
  if($line_mrna =~ /^>/)
  {
    $header = $line_mrna
  }
  unless($line_mrna =~ /^>/)
  {
    $seq_mrna = join('',$seq_mrna,$line_mrna);
  }
}
$seq_mrna =~ s/[\n\r]//g;   #string containing mRNA sequence.

#extract seed sequence of miRNA and search for it in mRNA sequence.

$header =~ s/>//;
print RESULT
      "The following miRNAs have $seed_type matches to:\n$header \n";

$num_mirna = 0;
```

```
  for $line_mir (<MIRNAS>)
  {
    if ($line_mir =~ /^>/)
    {
      $identifier = $line_mir;
      $identifier =~ s/[\n\r]//g;
      $identifier =~ s/>//g;

      @splitted_id = split(/ /,$identifier,2);
      $mirna_name = @splitted_id[0];
    }
    else
    {
      $seed = substr($line_mir,$seed_start,$seed_length);
      $seed_compl = reverse(generateComplementary($seed));

      if($A_anchor == 1)
      {
        $site = join('',$seed_compl,'A');
      }
      else
      {
        $site = $seed_compl;
      }

      @matches = searchForSequence($seq_mrna,$site);

      if (@matches != ())
      {
        $positions = join("\t",@matches);
        print RESULT "$mirna_name\t$positions\n";
        $num_mirna++;
      }
    }
  }
  if ($num_mirna == 0)
  {
    print RESULT "No matches for this mRNA where found.\n";
    print "No matches for this mRNA where found.\n";
  }
  else
  {
    print RESULT "$num_mirna miRNAs target this mRNA.\n";
    print "$num_mirna miRNAs target this mRNA.\n";
  }
}

#-------------------------------------------------------------------------
#The sub "generateComplementary" returns the complementary sequence
#to a given sequence.

sub generateComplementary
{
  my (@parameter) = @_;

  @sequence = split(//,$parameter[0]);

  for $nucleotide (@sequence)
  {
    if ($nucleotide eq 'A') {
      $nucleotide = 'T';
    } elsif ($nucleotide eq 'U') {
```

```
      $nucleotide = 'A';
    } elsif ($nucleotide eq 'T') {
      $nucleotide = 'A';
    } elsif ($nucleotide eq 'G') {
      $nucleotide = 'C';
    } elsif ($nucleotide eq 'C') {
      $nucleotide = 'G';
    }
  }
  return join('',@sequence);
}

#-------------------------------------------------------------------------
#The sub "searchForSequence" returns the positions in a given sequence,
#where a second seqences was found.

sub searchForSequence
{
  my (@parameter) = @_;

  my $seq = $parameter[0];
  my $compl_seq = $parameter[1];

  @position = ();

  while($seq =~ /$compl_seq/g)
  {
    $site_pos = pos($seq);
    @position = (@position, $site_pos);
  }
  return @position;
}
```