Biochem 218 Final Project
Hye Young Min (5360242)
hymin@stanford.edu

A review of computational methodsfor miRNA target prediction

Introduction

MicroRNAs (miRNAs) are a class of small, non-coding regulatory RNAs that are important in post-transcriptional gene silencing (Bartel et al, 2004). They regulate gene expression by binding to 3' untranslated region (UTR) of their target mRNAs for cleavage or translational repression and play important roles in many biological processes including cell proliferation, cell death, hematopoiesis, and oncogenesis.

In the canonical pathway of miRNA biogenesis, mature miRNAs arise from long primary miRNA transcripts (pri-miRNAs) that are transcribed from non–protein-coding genes in the nucleus (Figure 1; Lodish et al, 2008). The pri-miRNAs are then cleaved by the RNase III enzyme Drosha to liberate ~ 70-nt precursor miRNAs (pre-mRNAs) which are subsequently transported into the cytoplasm by Exportin-5, a Ran–GTP-dependent nuclear export factor. In the cytoplasm, the pre-miRNAs are processed by RNase III–like nuclease Dicer (animals) or DICER-LIKE1 [DCL1 (plants)] to generate ~21 to 22 nt duplexes. The functional mature miRNA strand is then selectively incorporated into RISC (RNA-induced silencing complex) effector complex to regulate specific target mRNAs. In general, plant miRNAs interact with their targets through near-perfect base-paring, resulting in target degradation, whereas animal miRNAs form imprecise base-pairing and cause translational repression.

Since the discovery of the very first miRNAs, computational approaches have been invaluable tools in understanding the biology of miRNAs (Bentwich et al., 2005; Rajewsky et al., 2006). Web-based-miRNA databases have been constructed and provided not only thousands of published miRNA sequences and annotation (miRBase Sequences) but also potential miRNA target genes (miRBase Targets). Many primary miRNA transcripts (pri-miRNAs) are computationally predicted to undergo folding into elaborate stem-loop structures. In addition, computer algorithms are developed to predict

pre-miRNAs (Huang et al., 2007) and to search for homologous conserved miRNA genes in several animal species. However, most computational approaches associated with miRNA research are miRNA gene detection and miRNA target prediction.

Researchers initially determined miRNA targets through experiments. The first miRNAs and their target genes had been identified through classical genetic techniques (Lee et al., 1993). However, due to the laborious nature of experiments and the absence of high-throughput experimental methods, it is inevitable to develop computational techniques to determine miRNA targets. In this paper, I summarize the principles to predict miRNAs and their targets, and discuss the currently available computational methods that have been developed for miRNA targets prediction.

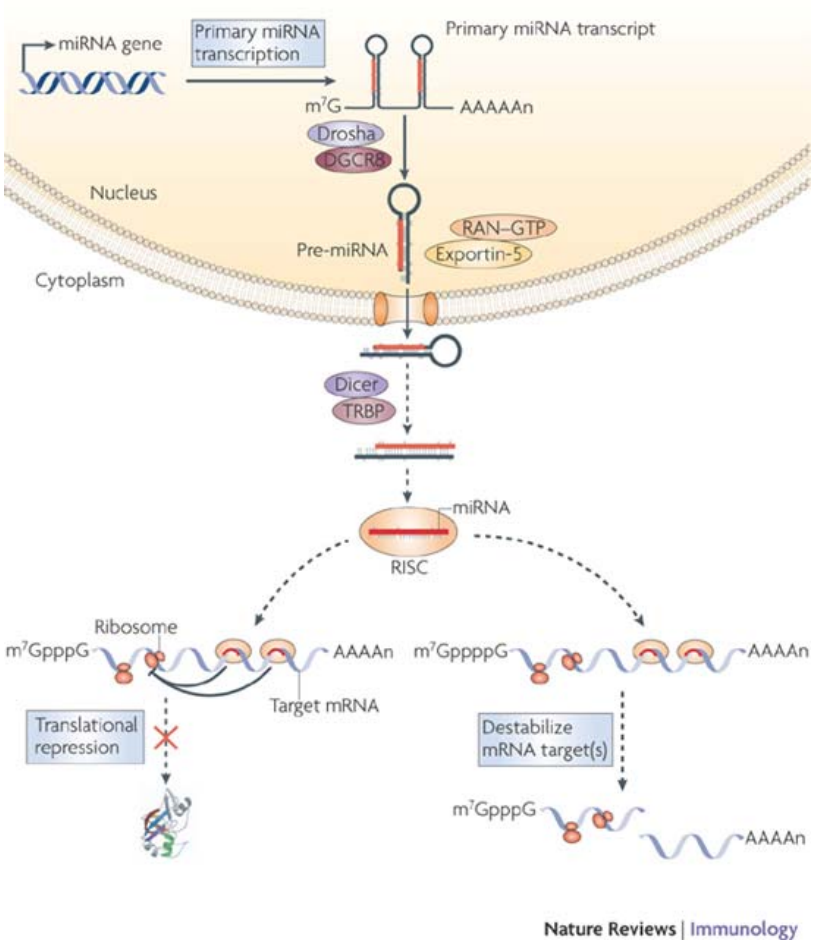

Figure 1. MicroRNA biogenesis and function in animal cells
(Adapted from *Lodish et al.*, 2008)

Principles of miRNA target recognition

Target prediction and its biological validation have been major obstacles to miRNA researcher. Because miRNAs are short, and animal miRNAs have limited sequence complementarity to their targets, it is a challenging task to predict animal miRNA targets with high specificity. For plants, target prediction is rather straightforward since plant miRNAs are believed to base pair to their targets with perfect or nearly perfect complementarity.

In order to develop computational algorithms identifying miRNA target genes, principles of miRNA target recognition are often established based on empirical evidences. For example, the importance of base pairing between miRNAs and their targets has been suspected according to the observation that the 'target site' of the lin-14 UTR is complementary to the 5' region of the lin-4 miRNA (Lee et al., 1993). Some features used by the mammalian target prediction programs are described below.

1) Base pairing pattern
2) Thermodynamic stability of miRNA-mRNA duplex
3) Comparative sequence analysis to check conservation
4) Checking for the presence of multiple target sites

Base pairing pattern

In the first step, target prediction programs identify potential binding sites according to specific pairing patterns. The binding sites can be classified into 3 categories (Maziere et al, 2007): (i) 5'-dominant canonical, (ii) 5'-dominant seed only and (iii) 3'-compensatory (Figure 2). MiRNA seed is defined as the consecutive 7 to 8 nt sequence starting from either the first or second base at the 5' end of an miRNA (Lewis et al, 2003). The 5' -dominant canonical sites have perfect base paring to the 5' end seed region and extensive base pairing to the 3' end of the miRNA with a characteristic bulge in the middle. The seed only sites have perfect base pairing to the seed region and limited base pairing to the 3' end of the miRNA. The 3'-compensatory sites have a mismatch or wobble in the seed region of the miRNA, but have extensive base pairing to the 3' end of the miRNA to compensate for the weak binding at the 5' seed (Brennecke et al., 2005).

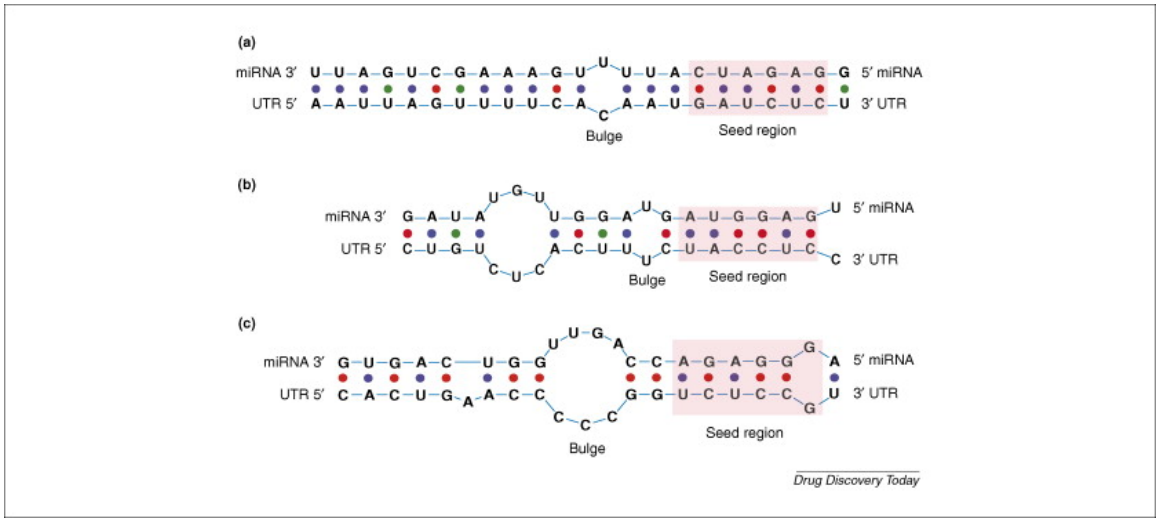

Figure 2. Approximate secondary structures of the three main types of target site duplex. (a) Canonical sites (b) Dominant seed sites (c) Compensatory sites
(Adapted from *Maziere et al.*, 2007)

Thermodynamic analysis of miRNA–mRNA duplexes

The thermodynamic properties of miRNA–mRNA duplex formation are analyzed by calculation of free energy (ΔG). The estimate free energy and secondary structure of the miRNA-mRNA duplex can be calculated by RNA folding program such as Vienna package (Wuchty et al., 1999). The free energy threshold is then calculated based on both specificity and sensitivity. However, it is very difficult to determine the appropriate thresholds of free energy, because data sets of known miRNA–mRNA duplexes are very limited, and a lower free energy (stable binding) does not always result in reliable prediction of miRNA target genes. Thus, it is necessary to consider other characteristics such as conservation analysis as discussed below. For example, a recent study by Lewis et al. (2005) has shown that thermodynamics can be removed without lowering the specificity of the algorithm by incorporating evolutionary conservation as an informational filter.

Comparative sequence analysis

Cross-species sequence comparison is used to ask whether the target sequence has been evolutionarily conserved between related species. In order to reduce the number of false positives, many of the target prediction algorithms identify orthologous 3’ UTR

sequences and then perform conservation analysis across species. However, there are issues of using conservation analysis. For instance, given that transcripts between humans and chimpanzees are highly conserved, it might not be meaningful to search for conserved targets between humans and chimpanzees (Maziere et al. 2007). Other species such as rats and dogs might be more relevant for comparing with human transcripts, but the fact is that genomes are not sequenced according to their evolutionary distance. As a result, the use of conservation filter has a risk of increasing false negatives.

Checking for the presence of multiple target sites

Previous studies have shown that mRNAs are likely to be regulated by multiple miRNAs. Multiple target sites in the same 3' UTR can potentially increase the degree of translational suppression. Thus, some algorithms count the number of target sites and check for the presence of multiple target sites.

Programs for miRNA target recognition

Different methods have been developed for computational target prediction. Currently available target prediction programs are listed in Tables 1, and some of them are reviewed below in more detail.

TargetScan /TargetScanS

TargetScan is an algorithm developed by Lewis et al. (2003) to identify the targets of vertebrate miRNA. The program combines thermodynamics-based modeling of RNA-RNA duplex interactions with comparative sequence analysis to predict miRNA targets conserved across multiple genomes such as human, mouse, rat, and pufferfish.

The ‘miRNA seed’ is a 7 nt sequence at base 2 to 8 in the 5’ end of the miRNAs. It forms perfect Watson-Crick complementary to ‘seed matches’ which refers to 3’ UTR heptamer in the target mRNA. TargetScan searches for seed matches in the first organism such as human and extends each seed match with additional base pairs to the miRNA. The algorithm then calculates the thermodynamic free energy of the binding between the putative miRNA target and extended seed sequence by using the RNAFold package (Hofacker, 2003) and assigns a score to each UTR. Then, it repeats the process for the

sets of UTRs from other organism including mouse rat, and pufferfish for phylogenic analysis. The estimated false-positive rate varies between 22 % and 31%, and the method was shown to predict not only known miRNA binding sites but also 451 novel potential sites. In addition, by using luciferase reporter constructs, 11 out of the 15 tested sites were experimentally validated.

TargetScanS simplified the TargetScan method and improved the target prediction fidelity (Lewis et al., 2005). TargetScanS requires a six-nucleotide seed (position 2 to 7) followed by an additional 3' match of adenosines surrounding the miRNA seed (It was found that the immediate downstream position of the seed match is highly conserved and is often an adenosine). The method is independent of thermodynamic stability or multiple target sites, but two more species (dog and chicken) were added for conservation analysis. As a result, estimated false-positive rate was reduced to 22% in mammals, and all known miRNA-target interactions were successfully predicted.

Although the TargetScan and TargetScanS efficiently reduced false positive rates, there is a concern about using conservation analysis and complementarity in the seed region. As shown in Figure 2 (c), 3' compensatory site has a mismatch or wobble in the seed region and does not form a perfect Watson-Crick base pairing. Therefore, some targets having 3' compensatory site cannot be detected. In addition, as mentioned earlier, if targets are loosely conserved, they will not be picked by TargetScan/TargetScanS resulting in an increase of false negatives.

PicTar

Contrary to TargetScan/TargetScanS that requires a seed match at exactly corresponding positions in a cross-species UTR alignment, PicTar requires only that the seed match occurs at overlapping position in a cross-species UTR alignment (Grun et al., 2005). This algorithm scans the alignments of 3'UTRs for those displaying seed matches to miRNAs. The retained alignments are then filtered according to their thermodynamic stability. PicTar then computes a Hidden Markov Model (HMM) maximum likelihood score (PicTar score) that a given RNA sequence (typically a 3' UTR) is targeted by combinations of microRNAs. This algorithm was able to correctly identify known miRNA targets and its false-positive rate has been estimated to be around 30%.

By using PicTar, Krek et al. suggested that each vertebrate miRNA targets approximately 200 transcripts on average. In addition, they experimentally validated 7 out of 13 predicted targets and 8 out of 9 previously known targets, demonstrating the efficiency of the algorithm. Furthermore, Grun et al. (2005) exploited cross-species comparison to predict that on average, 54 genes are regulated by a given miRNA. PicTar was also applied to genome-wide search of miRNA targets in *C. elegans* (Lall et al., 2006). By using a new version of PicTar and sequence alignments of three nematodes, the authors reported that at least 10% of *C. elegans* genes are predicted miRNA targets, and a number of nematode miRNAs seem to regulate biological processes by targeting functionally related genes.

miRanda

This method was originally developed to predict miRNA target genes in *D. melanogaster* (Enright et al., 2003), but also used to predict human miRNA targets. For each miRNA, miRanda selected target genes on the basis of three properties: sequence complementarity using a position-weighted local alignment algorithm, free energies of RNA-RNA duplexes, and conservation of target sites in related genomes. The method correctly identified 9 of 10 currently characterized target genes, and its false-positive rate was estimated to be 24%. When analyzed the distribution of functional annotation for all targets of all miRNAs using Gene Ontology (GO) terns, the functions of the predicted target genes were enriched in the components of the ubiquitin machinery, transcription factors, components of miRNA machinery, and translational regulation.

John et al. (2004) improved the method by implementing a strict model for the binding sites that requires almost perfect complementarity in the seed region allowing a single wobble pairing. The authors reported about 2000 human genes with miRNA target sites conserved in mammals and about 250 human genes conserved between mammals and fish. Their analysis also suggests that miRNA genes, which are about 1% of all human genes, regulate protein production for 10% or more of all human genes.

DIANA-microT

Kiriakidou et al. (2004) developed this method by combining computational and experimental approaches. In order to identify putative miRNA-recognition elements (MREs), this method uses a window of 38-nt that is progressively moved across a 3′UTR mRNA sequence. Using dynamic programming, the minimum binding energy between the miRNAs and sequences in the human 3'-UTR database is calculated at each step and compared with the results obtained from shuffled sequences with the same dinucleotide composition as real 3′UTRs.

In contrast to TargetScan/TargetScanS or PicTar, this method allows a central bulge and strong binding at 3' end of miRNA when 5' seed pairing is rather weak. In addition, unlike the previous works, this method uncovers predominant miRNA targets that contain only single target sites. This algorithm successfully identified all currently known *C. elegans* miRNA target sites. Moreover, 7 predicted mammalian miRNA target genes were experimentally validated.

RNAHybrid

RNAHybrid is an extension of classical RNA secondary structure prediction software tools such as RNAfold (Hofacker, 2003) and Mfold (Mathews et al., 1999). The classical methods were designed for single-sequence folding, and therefore require an artificial linker between the miRNA and its potential binding site. However, there are some issues about using the methods (Stark et al., 2003). The short artificial linker sequence might lead to artefacts in the prediction, and hybridizations of the target with itself, or of the miRNA with itself, or of both with the linker, can happen. An additional drawback is that the appropriate potential binding sites have to be cut out and folded separately for prediction of multiple bindings in one target, However, RNAHybrid finds the energetically most favorable hybridization sites of a small RNA within a large target RNA sequence, and base pairings between target nucleotides or between miRNA nucleotides are not allowed.

The method was successfully tested to predict known targets in *D. melanogaster* by using a 6-nt seed match starting from the second base of the 5' end of the miRNA. RNAHybrid has a low false-positive rate, and most of all, the association of P values with predicted targets is an appreciable asset for directly comparing predicted binding sites.

Discussion

Most computational algorithms for target prediction combine 5' seed matches, thermodynamic stability and conservation analysis in order to maximize specificity of the algorithms. However, there are some exceptions to these generalized rules, and it is also true that target selection mechanisms vary from species to species (Watanabe et al., 2007)

Although the rule of seed pairing has been successfully used to predict target sites with statistical support, the seed matches are not always sufficient for repression, indicating that other characteristics help specify targeting (Grimson et al., 2007). Through the combination of computational and experimental approaches, the authors revealed five general features of site context that boost site efficacy: AU-rich nucleotide composition near the site, proximity to sites for coexpressed miRNAs (which leads to cooperative action), proximity to residues pairing to miRNA nucleotides 13-16, positioning within the 3'UTR at least 15 nt from the stop codon, and positioning away from the center of long UTRs. Thus, in designing an algorithm, those five features as well as the rule of seed match should be considered.

Another problem of using 5' dominant site is that 3' compensatory site having a mismatch or wobble in the seed region cannot be detected by most target prediction methods. Among publicly available mammalian target prediction programs, the predictions provided by miRanda are the most sensitive for such targets (Sethupathy et al., 2006). Accordingly, it is of necessity to develop more computational algorithm to identify those 3' compensatory target sites with accuracy.

Evolutionary conservation is another important factor to filter out false positive targets and increase specificity. It helps to predict only the target sites which are under selective pressure to preserve their sequence, and presumably their functionality, across evolution (Sethupathy et al., 2006). However, Farh et al. (2005) demonstrated that many of the nonconserved target sites, which outnumber the conserved sites 10 to 1, are also functional and mediate repression. Thus, the presence of those nonconserved target sites should not be overlooked when designing an algorithm for target prediction.

Once miRNA targets are predicted with a fair degree of accuracy, the next step is to experimentally validate the miRNA – target interaction. Since computational methods are not perfect, and there is a risk of false-positive prediction, target validation in

biological system is inevitable to complete the study of target prediction. Reporter assay is the most common method to check the interaction between miRNA and its target mRNA. Then, Northern blot analysis, quantitative real-time PCR (qPCR), or *in situ* hybridization is often performed to examine the co-expression of predicted miRNA and mRNA target gene. For thorough study, biological function can be examined through 'gain of function' or 'loss of function' experiment under *in vitro* or *in vivo* condition. However, those biological or biochemical experiments (even the reporter assay) are laborious, time-consuming, and expensive to deal with many pairs of miRNAs and their targets. Therefore, high-throughput experimental strategies should be developed for large-scale analysis of miRNA targets and their biological function.

References

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

Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. FEBS Lett. 579, 5904–5910.

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

Chan, C. S., Elemento, O., and Tavazoie, S. (2005). Revealing posttranscriptional regulatory elements through network-level conservation. PLoS Comput. Biol. 1, e69.

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

Farh, K. K., Grimson A., Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science 310, 1817 - 1821

Gerlach, W., and Giegerich, R. (2006). GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. Bioinformatics 22, 762–764.

Grun, D., Wang, Y. L., Langenberger, D., Gunsalus, K. C., and Rajewsky, N. (2005). MicroRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Comput. Biol. 1, e13.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. 31, 3429–3431.

Huang, T. H., Fan B., Rothschild, M. F., Hu, Z. L., Li, K., and Zhao, S. H. (2007) miRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC bioinformatics. 8:341

Kim, S. K., Nam, J. W., Rhee, J. K., Lee, W. J., and Zhang, B. T. (2006). miTarget: microRNA target-gene prediction using a support vector machine. BMC Bioinform. 7, 411.

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

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, 495–500.

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y. L., Dewey, C. N., Sood, P., Colombo, T., Bray, N., Macmenamin, P., Kao, H. L., Gunsalus, K. C., Pachter, L., Piano, F., and Rajewsky, N. (2006). A genome-wide map of conserved microRNA targets in C. elegans. Curr. Biol. 16, 460–471.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843-854.

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 120,15–20.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. Cell 115, 787–798.

Lodish, H.F., Zhou, B., Liu, G., and Chen, C.Z. (2008). Micromanagement of the immune system by microRNAs, Nat. Rev. Immunol. 8, 120-130.

Maziere, P., and Enright, A. J. (2007). Prediction of microRNA targets. Drug Discov. Today. 12, 452 – 458.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940.

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell 126, 1203–1217.

Rajewsky, N. (2006). microRNA target predictions in animals. Nat. Genet. 38(Suppl.), S8–S13.

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

Saetrom, O., Snove, O., Jr., and Saetrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA 11, 995–1003.

Stark, A., Brennecke, J., Russell, R. B., and Cohen, S. M. (2003). Identification of Drosophila microRNA targets. PLoS Biol. 1, E60.

Watanabe, Y., Tomita, M., and Kanai, A. (2007). Computational methods for microRNA target prediction. Methods Enzymol. 427, 65-86.

Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49, 145–165.

Table 1. Computational methods for miRNA target prediction

Name	URL	Supported organism(s)	Reference
DIANA-microT	http://diana.pcbi.upenn.edu/cgi-bin/micro_t.cgi	Vertebrates	Kiriakidou *et al*., 2004
FastCompare	http://tavazoielab.princeton.edu/mirnas	Nematodes, flies	Chan *et al*., 2005
GUUGle	http://bibiserv.techfak.uni-bielefeld.de/guugle	Flies	Gerlach *et al*., 2006
miRanda	http://www.microrna.org/	Flies, vertebrates	Enright *et al*., 2003
miTarget	http://cbit.snu.ac.kr/miTarget	Any	Kim *et al*., 2006
PicTar	http://pictar.bio.nyu.edu	Nematodes, flies, vertebrates	Grun *et al*., 2005
rna22	http://cbcsrv.watson.ibm.com/rna22.html	Nematodes, flies, vertebrates	Miranda *et al*., 2006
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid	Flies	Rehmsmeier *et al*., 2004
TargetBoost	https://demo1.interagon.com/demo	Nematodes, flies	Saetrom *et al*., 2006
TargetScan	http://genes.mit.edu/targetscan/	Vertebrates	Lewis *et al*., 2003
TargetScanS	http://genes.mit.edu/targetscan/	Vertebrates	Lewis *et al*., 2005