Analysis of the Computational Methods Available for MicroRNA Target Prediction

Talayeh Ghezelayagh

Biochemistry 218: Computation Molecular Biology

MicroRNAs (miRNAs) are non-coding, endogenous RNAs that make up a vast post-transcriptional regulatory network in cells. Processed from primary transcripts that are translated from miRNA genes, mature miRNAs are usually 19-25 nucleotides long. More than 4000 miRNAs have been identified in the genomes of over 80 species. Mature miRNAs are thought to down-regulate the translation of messenger RNAs after recognizing and binding to partially complementary sites in the 3'-untranslated regions (3'-UTR) of the mRNA. MiRNAs have been found to be involved in the control of wide breadth of cell processes, including cell proliferation, death, and metabolism; therefore, their study is tremendously important for the understanding of cell function in eukaryotic species.[1]

The functional analysis of miRNAs relies heavily on the identification of their targeted genes. However, experimental studies alone are too slow and limited in scope to be relied on as the only source of miRNA target identification. In order to facilitate the investigation into miRNA function, numerous bioinformatic methods were developed in order to allow high-throughput prediction of miRNA target genes. In the last 7 years, numerous computational algorithms have been developed for target prediction spanning a wide range of approaches and techniques.[2] Several of these algorithms, namely miRanda, TargetScan, Pictar, TargetBoost, and PITA, will be discussed and analyzed in the following paper with a focus on possible areas of improvement (Table 1).

Table 1. Popular computational algorithms for microRNA target prediction

Software Name	*URL*	*Reference(s)*
TargetScan, TargetScanS	http://genes. mit.edu/targetscan/	Lewis et al., 2003, 2005
miRanda	http://www.microrna.org/ [or] http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/	Enright et al., 2003 John et al., 2004 Kiriakidou et al., 2004
Pictar	http://pictar.bio.nyu.edu/	Grun et al., 2005 Krek et al., 2005 Lall et al., 2006
TargetBoost	http://demo1.interagon.com/targetboost/	Saetrom et al., 2006
PITA	http://genie.weizmann.ac.il/pubs/mir07/index.html	Kertesz et al. 2007

Difficulties involved in miRNA target prediction

The prediction of target mRNAs for miRNAs in vertebrates is a particularly difficult challenge because of the complexity of miRNA target recognition. MiRNAs are short, typically only 19-25 nucleotides in length. Furthermore, miRNA:mRNA duplexes in vertebrates often contain several mismatches, gaps and G:U base pairs in many positions, which limit the maximum length of contiguous sequences of perfect nucleotide matching. This is in contrast to miRNA function in plants, where targets are usually recognized by perfect complementarity along the length of the miRNA.[3]

Specific base pairing patterns are found within miRNA:mRNA duplexes which must be accounted for in target recognition algorithms. In most cases, the 5' region of the miRNA perfectly complements the corresponding sequence of the target mRNA. This "seed region" is usually composed of 7 to 8 bases starting from either the first or second base of the 5' end.[2] [There is at least one report, however, where the target gene is regulated independently from base pairing in the seed region.][4] On the other hand, base pairing of the 3' region of miRNA is usually thought to be weaker and less important, unless stronger binding here compensates for

weaker binding in the seed region. Weaker binding is also usually observed at the central region of the miRNA:mRNA duplex, where mismatches cause bulge structures which may be useful for regulation of the mRNA. An example of a typical miRNA:mRNA duplex structure is shown in Figure 1.[2] Unfortunately, the specific and comprehensive rules controlling miRNA target recognition and binding have not been discovered,

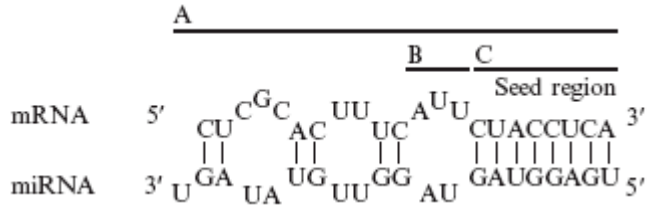

Figure 1. Example of a typical miRNA:mRNA duplex. Base pairing between *let-7a* miRNA and *hbl-1* mRNA in *C. elegans* demonstrates that base pairing is particularly weak in the central region and particularly strong at the 5' end on of the miRNA (seed region). Figure taken from Watanabe, Y. et al, *Methods in Enzymology*, 2007, 327: 65-86.

Common principles used in miRNA target prediction algorithms

Most miRNA target prediction algorithms use similar general principles in the development of their algorithm. Most algorithms search for targets in the 3'-UTR region of mRNAs, where almost all miRNA-target interactions occur in vertebrates.[5] Furthermore, algorithms usually account for the possibility of multiple target sites for more than one miRNA in each mRNA 3'-UTR region, though they differ in the degree in which combinations of miRNA target sites are incorporated into the prediction algorithm. Some algorithms take a "seed-based" approach, where prediction of binding sites heavily relies heavily on almost perfect complementation in the seed region, while others are less dependent on perfect complementarity within this region.[3]

Most miRNA target prediction algorithms, the likelihood that a particular miRNA:mRNA duplex will form *in vivo* is estimated by analyzing the thermodynamic properties of the duplex structure. The Vienna package is the most commonly used software tool for the estimation of free energy and RNA secondary structure.[6] The various algorithms set free energy thresholds to evaluate possible miRNA-target interactions. MiRNA target prediction algorithms often make

use of the RNAfold program, which predicts minimum energy secondary structures and pair probabilities and the RNAduplex program, which predicts possible hybridization sites between two RNA sequences. The software can be accessed at http://www.tbi.univie.ac.at/RNA/.[7]

The degree of sequence conservation is another criterion commonly used to filter possible miRNA targets. Most target prediction algorithms identify orthologous 3'-UTR sequences and check whether the miRNA-target interaction is conserved between closely related species.[2] If miRNAs are conserved in orthologous species, their targets are also expected to be conserved, therefore allowing for a reduction in false positive predictions and a useful filter to use since perfect miRNA binding rules are as-of-yet unknown.

Many algorithms depend on an initial input of a miRNA or miRNAs to be queried and a set of genes to test for targets. Researchers draw information for their analysis from useful databases. The miRNA registry, or miRBase, is a widely used database of miRNA sequences.[8,9] A web interface (http://microrna.sanger.ac.uk/) allows users to search and browse for miRNAs in multiple species and allows for the downloading of sequence information. Sequences of mRNAs, particularly of their 3'-UTR, are often obtained from the Ensembl database, which allows users to generate the 3'-UTR sequences for all transcripts of all genes from each genome.[10]

miRanda

The miRanda algorithm, introduced in 2003, was one of the first miRNA target prediction algorithms to be developed and is now one of the most heavily used.[11] It is now used for target prediction by multiple interfaces accessible to researchers, including http://microRNA.org[12] and MicroCosm Targets, accessible at http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/. The algorithm ranks the likelihood of each gene to be a miRNA target and the likelihood of each miRNA to target a gene. Overall, the algorithm relies on a stepwise progression through

target prediction, free-energy calculation to estimate the thermodynamics of the possible interactions, and evolutionary conservation.

The algorithm first depends on the retrieval of miRNA and 3'-UTR sequences. Given this set, the target prediction component of the algorithm consists of using dynamic programming to search for maximal local complementarity alignments between the mature microRNA sequence and all possible positions in each gene 3'-UTR. A score of +5 is assigned for G:C and A:T pairs, +2 for G:U wobble pairs, and -3 for all other nucleotide pairs. The opening of a gap has a penalty of -8 while gap elongation costs -2. The gap-elongation parameter was only used if the cost of a gap at a position is less than the cost of a mismatched base pair. The miRanda algorithm recognizes the importance of the seed region by multiplying the complementarity scores (positive and negative values) at the first eleven positions, from the miRNA 5' end, by a scaling factor of 2. The value of the scaling factor is an adjustable parameter that can be optimized as new experimental information is discovered about the seed region. Furthermore, four empirical rules are applied in order to ensure that the proposed miRNA:mRNA duplex follows experimentally determined patterns: no mismatches at positions 2 to 4 (counting from the 5' end); fewer than five mismatches between positions 3-12; at least one mismatch between positions 9 and L-5 (L is the total alignment length); and fewer than two mismatches in the last five positions of the alignment. The dynamic programming algorithm thus optimizes the complementarity score between a miRNA sequence and mRNA sequence summed over all aligned positions and creates a ranking of all non-overlapping hybridization alignments in decreasing order of score down to some cut-off value (cutoff typically 80 or 90). The key extension of the miRanda algorithm over the Smith-Waterman sequence alignment algorithm is the addition of weighted scores for certain positions in the alignment.[11-13]

To estimate the stability of the predicted miRNA:mRNA duplex that would be created from each alignment, the miRanda algorithm uses folding routines from the Vienna package. The free energy of the duplex is estimated and checked against a threshold value. Previous

implementations of this algorithm have used G<-17 kcal/mol or G<0-14 kcal/mol as cut-off values.[11-13]

Finally, the miRanda algorithm searches for conservation of the target site between orthologous species. The species used in this analysis obviously depends on the reference species. For example, the initial prediction of targets in the human genome consisted of a search for conservation between human, mouse, and rat.[13] The alignments of target sites are generated transitively via a shared homologous miRNA (i.e. UTR target in human -> miRNA -> UTR target in mouse). The positions of pairs of target sites in to species must fall within ±10 residues in the aligned 3' UTRs. The sequence identity between conserved target sites must also meet a certain threshold (≥90% used for analysis in humans[13]).

After passing the conservation filter, predicted target sites for each miRNA are sorted according to alignment score primarily and free energy secondarily. Overall, however, the miRanda algorithm is a very popular tool for miRNA target prediction. It was used initially to predict miRNA target genes in *D. melanogaster*, correctly identifying 9 of 10 published miRNA-target interactions, with a false positive rate of 24%.[11] However, this algorithm does have flaws that can be improved upon. If multiple miRNAs target the same site on a transcript, only the highest scoring, lowest energy miRNA is reported for that site. This parameter is a possible source of false negatives because different miRNAs are expressed at different times in cell development.[1] It is possible that multiple miRNAs can bind to overlapping sites, but because they are never expressed at the same time in the cell, they don't usually compete for the binding. Furthermore, Ioshikhes *et al*. argues for less reliance on Smith-Waterman alignment in algorithms such as miRanda.[14] The Smith-Waterman alignment was originally designed for comparing evolutionarily related sequences and miRNAs and their mRNA targets do not fall into this category. A better alignment algorithm that is specifically designed for miRNA/mRNA sequence comparison will improve the miRanda algorithm. Finally, the values and penalties

assigned throughout the algorithm are arbitrary and should be modified as greater knowledge about the biology of miRNA binding accumulates.

TargetScan/TargetScanS

The TargetScan algorithm places a greater emphasis on the seed region than miRanda but follows many of the same general principles.[15] To predict targets in a queried animal, the algorithm begins with a miRNA that is conserved in multiple organisms and a set of orthologous 3'-UTR sequences from the genes in these organisms. TargetScan then searches the UTR regions of the queried species for segments of perfect Watson-Crick complementarity (G:C or A:U) to bases 2 to 8 of the miRNA, numbered from the 5' end (termed the "seed match"). The algorithm then extends the seed match with additional base pairs to the miRNA as far as possible in each direction, allowing G:U wobble pairs, but stopping at mismatches. After stopping, the algorithm uses the RNAfold program of the Vienna package to optimize base pairing of the remaining 3' portion of the miRNA to the 35 bases of the UTR after the seed match. The resulting predicted miRNA-target interaction is then assigned a folding free energy.[15]

A final score Z is used to rank the likelihood that the UTR is a target for the miRNA: $Z = \sum_{k=1}^{n} e^{-G_K / T}$, where n is the number of seed matches in the UTR, G_K is the free energy of the miRNA:mRNA duplex for the k^{th} target site. T is a parameter that influences the relative weighting of UTRs with fewer high-affinity sites to those with larger numbers of low-affinity sites. The value of T was determined through testing to be optimized at a value of 10 (for an organism set involving human, mouse, and rat). The UTRs in the organism is sorted by Z score and then assigned a rank R_i. This process is repeated for the set of UTRs in each organism in the set. The final list of predicted targets includes those where both $Z_i \geq Z_C$ and $R_i \leq R_C$ for an orthologous UTR sequence in each organism in the set. Like T, the Z_C and R_C cut-off values

are determined through optimization and have previously been set as 4.5 and 350, respectively. The TargetScan Algorithm thus relies heavily on complementarity in the seed region, free energy calculations, and conservation in orthologous UTRs.[15]

Two years after the initial algorithm was created, changes were made to create the TargetScanS algorithm.[16] Additional whole genomes (chicken and dog) were added to the conservation set, which reduced the estimated number of false-positive predictions and allowed for the total elimination of Z_C and R_C cut-offs in favor of dependence almost solely on seed matches. The conservation criterion was also extended to require that the conserved seed matches be at conserved positions within the UTRs, like the miRanda algorithm. Furthermore, the seed region was reduced by one base to require perfect complementarity only between nucleotides 2 through 7. Finally, since experimental evidence showed bias towards the presence of certain nucleotides in key positions, the TargetScanS algorithm predicts targets that have a conserved 6-nt seed match flanked by either a Watson-Crick match at position 8 of the miRNA or an adenosine immediately downstream of the seed region in the target sequence.[16]

Initial testing of the TargetScan algorithm resulted in experimental validation of 11 out of 15 predicted targets in mammalian cells. The fraction of false positive results was estimated at 31% for targets identified in human, mouse, and rat. However, out of over 5300 human genes predicted as potential targets of miRNAs, a significant number of genes are correctly identified as miRNA targets to be used as candidates for further experimental validation.[15] The advantages of TargetScanS include reduced false positive predictions because of the heavy emphasis on conservation.[16] However, these algorithms may be limited in its applicability since predictions are restricted to miRNAs conserved between the species with less than one substitution. MiRNAs that are conserved but may have additional substitutions (perhaps in non-essential locations) cannot be tested. Furthermore, both algorithms are heavily dependent on perfect complementarity in the seed region, therefore missing targets with G:U wobble pairs in the seed region.[2]

PicTar

The advantage of the PicTar (probabilistic identification of combinations of target sites) algorithm is its recognition that the degree of translational repression may increase exponentially with the number of miRNA binding sites in the 3'-UTR, therefore it is important to search for combinations of miRNA binding sites for sets of co-expressed miRNAs.[17] PicTar depends on the input of a fixed search set of miRNAs and multiple alignments of orthologous nucleate sequences (3'-UTRs usually) and outputs scores that rank genes by their likelihood of being a common target of subsets of the miRNA search set and probabilities for the predicted binding sites in each UTR. To accomplish this, PicTar uses hidden Markov models to model the 3'-UTR of a gene where the HMM states are bindings sites of each of the miRNAs from the set and the background.

The miRNA binding sites are represented by 7-8 nucleotide long "nuclei", (mRNA stretches Watson-Crick base paired from the first or second position from the 5' end of the miRNA). Insertions or mutations in the mRNA sequence of a perfect nucleus is allowed if the free energy of binding does not increase and does not contain G:U wobble pairs. The free energy of the miRNA:mRNA duplex must also be below a cutoff value (33% of the optimal free energy of the entire mature miRNA binding to a perfectly complementary target site for sites with perfect nuclei; 66% of the optimal free energy for sites with imperfect nuclei). Nuclei that survive this filter are assigned a probability p to be a binding site for the miRNA (p usually assigned to be ≈0.8 for perfect nuclei and (1-p)/(# of imperfect nuclei) for imperfect nuclei).[17] The PicTar program first looks for the conserved 3' UTR segments containing minimal numbers of perfect and imperfect matches for a given miRNA set specified by the user. These segments are used to derive a HMM-based score for a given UTR to be targeted by the given miRNA.[14]

PicTar was initially used to predict miRNA targets in vertebrates. Testing showed that it had the ability to recover published miRNA targets and has a false positive rate around 30%. Advantages of this algorithm include ability to output probabilities of binding and its use of

combinatorial methods to account for binding by multiple miRNAs. However, it still relies on the same sequence alignment and conservation analysis done in other algorithms.[2]

TargetBoost

In contrast to the algorithms described previously, TargetBoost takes a machine-learning approach to miRNA target prediction by combining genetic programming with boosting.[18] The algorithm essentially tries to learn the hidden rules of miRNA-target site hybridization without relying on criteria based on sequence complementarity, thermodynamic stability, or evolutionary conservation. Given a miRNA and a potential target site, TargetBoost reports a score that represents the likelihood of the site being targeted by the miRNA.

The TargetBoost algorithm relies on training from a set of known miRNA targets (and negative controls). From this set, the program creates a sequence-based classifier that recognizes the positive sequences. The classifier is a sum of several differentially weighted "template queries", which are general expressions that describe the common properties of miRNA target sites. These template queries are translated into queries that are specific for each miRNA in the search. The individual template queries are developed using a combination of genetic programming[19] to evolve the individual queries from a population of candidate patterns and boosting[20] to guide the search by adjusting the importance of each sequence in the training set. The boosting algorithm then assigns weights to the template queries based on their individual performance in the training set. The template queries are then used to search a set of 3'-UTR sequences for binding sites to a desired miRNA.[18]

This algorithm thus develops a set of rules based on a training set, and applies those rules in the search of new miRNA binding sites. When the algorithm was tested using 10-fold and leave-one-miRNA-out cross-validation (use 9 miRNA training sets and check predictions for the 10th one), TargetBoost was found to have a high ROC_{50}-score, which is the area under the ROC curve until 50 false positives are found, of 0.0025, performing better than other prediction

algorithms RNAhybrid and Nucleus (both not described here). TargetBoost also confirmed the tendency of perfect matching in the 5' end of the miRNA.[18]

A potential flaw of TargetBoost is that it treats G:U wobbles as normal mismatches, and thus would not find potential target sites with a high number of G:U wobbles. Indeed, TargetBoost did not recognize the miR-92a target in *tailless* and the miR-210 target in *hairy* that have been validated in *D. melanogaster*.[18] However, because TargetBoost relies on more than thermodynamic stability and binding in the seed region through its development of specific classifiers, better classification and prediction can be achieved. The programming and boosting algorithms may be able to identify additional patterns in miRNA-target binding that are not frequently reported and thus not incorporated into other prediction algorithms. Finally, though the initial applications of TargetBoost did not do so, additional filters such as requiring conservation of the target sites or the presence of multiple target sites in the 3'-UTR can be added to reduce the number of false positive predictions.

PITA

The PITA algorithm was one of the first to incorporate target accessibility into miRNA target site prediction.[19] Because there is an energetic cost to freeing base-pairing interactions within mRNA secondary structure, the creators of this algorithm believed that secondary structure thus contributes to target recognition and is considered in the analysis. The PITA algorithm involves first scanning the 3'-UTR for perfect complementarity to miRNA seed regions (at least seven bases long) , and then applies a thermodynamic model to each such putative site, finally combining sites for the same miRNA to obtain a total interaction for the miRNA and UTR. The search can be limited to a set of evolutionary conserved miRNAs and mRNAs if desired.

The thermodynamic model scores microRNA-target interactions by an energy score $\Delta\Delta G$, which is equal to the difference between the energy gained by binding of the miRNA to

the target, ΔG_{duplex}, and the energy required to make the target region accessible for miRNA binding, ΔG_{open}. To compute the ΔG_{duplex} value, the RNAduplex program in the Vienna Package is modified to predict hybridization structure given explicit seed pairings. The program then computes the binding free energy of the complying structures and selects the minimum free energy structure as ΔG_{duplex}. ΔG_{open} is calculated as the difference between the free energy of the ensemble of all secondary structures of the target region and the free energy of all target-region structures in which the target nucleotides (and additional nucleotides upstream and downstream in the case of flanking) are required to be unpaired. Using dynamic programming, the free energies of these two ensembles are computed using RNAFold, the Vienna package program that predicts minimum energy secondary structures, by iterating over all possible structures and summing their free energies. RNAFold is given the area of the target and 70 additional nucleotides upstream and downstream to analyze for secondary structure (distances greater than 70 nucleotides away have low probabilities of structural interactions). Therefore, this value estimates the energy required to unpair the nucleotides that will be involved in recognition and binding of the miRNA. In order to integrate multiple sites with $\Delta\Delta G$ scores for a single miRNA on the same UTR, the algorithm computes the overall miRNA-UTR interaction score T using the formula: $T = -\log\left(\sum_{i=1}^{n} e^{-\Delta\Delta G_i}\right)$.[19]

The score T is calculated to represent the configuration that only one of the sites is bound at any given time. The algorithm thus does not account for the possibility in which two or more sites can be bound simultaneously.[19] Such a calculation would require knowledge of the free concentration of the miRNA, which is unknown as-of-yet, but would vastly improve the biological relevance of the search. However, development of the PITA algorithm did improve the standard thermodynamic analysis conducted in previous target prediction algorithms.

Comparing the available algorithms for miRNA target prediction

Six different algorithms were presented here that span the majority of computational approaches used for miRNA target predictions. MiRanda uses weighted target prediction scores, thermodynamics, and conservation analysis. TargetScan relies on perfectly-matched seed regions along with thermodynamics and conservation analysis, while TargetScanS relies on seed region complementarity and high conservation. PicTar uses combinatorial methods to predict miRNA binding in the presence and activity of other miRNAs. TargetBoost takes a machine-learning approach to generate miRNA-target interaction rules used to predict further targets. The PITA algorithm takes into account target site accessibility by factoring mRNA secondary structure before miRNA binding. Each algorithm has its own advantages and limitations, therefore researchers can pick the algorithm that best fits their needs.

A quantitative comparison of these algorithms is difficult to conduct because no study has been down using all these algorithms for the prediction of the same targets. However some qualitative comparisons can be made due to their general approaches. miRanda, TargetScan, PicTar and PITA are all limited by their reliance on currently established miRNA recognition and binding rules. While the free energy of the miRNA-duplex is taken into account, emphasis is profoundly placed on seed region matches. Evolutionary conservation is used as a replacement for more detailed recognition and binding rules that are not yet discovered. More subtle requirements (or biases) for miRNA-target interactions that are not utilized for prediction yet (or perhaps not even identified) can vastly improve the specificity of miRNA target prediction. Machine learning approaches such as TargetBoost have an advantage in this area because they will continue to discover and incorporate these subtle rules as more and more miRNA-interactions are validated and added to training sets. However, there is danger that the use of computer-calculated rules will limit the sensitivity of predictions. MiRanda, which does not rely as heavily on seed matches as TargetScan, PicTar, and PITA, in favor of user-defined cut-off scores, may provide the greatest number of predictions for miRNA targets. This is useful for

studies desiring high sensitivity rates. Groups studying the binding and function of co-expressed miRNAs may benefit from PicTar's combinatorial model, however, while TargetScan and PITA may be preferred for other reasons.

Limitations to current algorithms and suggestions for improvement

While the algorithms presented here and elsewhere are useful for miRNA target prediction and have been embraced by the scientific community, improvements can still be made to more precisely predict targets with less false positive and false negative results. It is interesting to consider ways in which the current algorithms can learn from each other. For example, machine-learning approaches like TargetBoost could be improved through the addition of checkpoints testing target site availability as used in the PITA algorithm. A combination of the miRanda and PITA algorithms could result in less reliance on seed matches but greater knowledge of target availability. The advantages of multiple algorithms can be combined in order to produce more accurate target predictions. Greater cooperation and communication within the bioinformatics research groups tackling this problem may improve target prediction in the future.

Currently, the application of computational miRNA prediction algorithms has mainly involved searching for targets with the 3'-UTR. However, it has been shown that target sequences inserted in the coding or 5'-UTR regions can also be functional.[22-23] A greater emphasis should be placed on searching the entire gene (5'-UTR, introns, exons, and 3'-UTR) for potential miRNA targets in order to reduce the number of false negatives in all prediction algorithms. Furthermore, the algorithms that make use of evolutionary conservation should be relaxed in order to better fit the needs of more research projects. If the incorporation of conservation analysis was made optional, prediction algorithms can be used on miRNAs and genes specific to one species, or for those whose homologues have not yet been identified in other species.

Ultimately, our knowledge of miRNA recognition and binding patterns remains incomplete. Computational methods are thus limited in their accuracy and scope until increased research in the field is conducted. It has recently been reported that nucleotides in the loop region of precursor miRNAs affect miRNA activity (and thus, perhaps affects binding), which questions the current dependence on mature miRNA sequence.[24] If this finding is further validated, all the target prediction algorithms currently used will need to be updated. At the moment, however, the algorithms in place are effectively predicting new candidate miRNA-target interactions for use in defining the roles and functions of miRNAs in organisms.

References

1. Bartel, D.P. "MicroRNAs: genomics, biogenesis, mechanism, and function." *Cell*, 2004, 116(2): 281-297.
2. Watanabe, Y. et al. "Computational methods for microRNA Target prediction." *Methods in Enzymology*, 2007, 427: 65-85.
3. Mendes, N.D. et al. "Current tools for the identification of miRNA genes and their targets." *Nucleic Acids Research*, 2009, 37(8): 2419-2433.
4. Didiano, D. and Hobert, O. "Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions." *Nature Structural & Molecular Biology*, 2006, 13: 849-851.
5. Rajewsky, N. "microRNA target predictions in animals." *Nature Genetics*, 2006, 38: S9-S18.
6. Hofacker, I.L. "Fast folding and comparison of RNA secondary structures." *Monatshefte fur Chemie*, 1994, 125(2): 167-188.
7. Hofacker, I.L. "Vienna RNA secondary structure server." *Nucleic Acids Research*, 2003, 31(13): 3429-3431.
8. Griffiths-Jones, S.R.B. "The microRNA sequence database." *Methods in Molecular Biology*, 2006, 2006, 342: 129-138.

9. Griffiths-Jones, S.R.B et al. “miRBase: microRNA sequences, targets, and gene nomenclature.” *Nucleic Acids Research*, 2006, 34: D140-144.
10. Hubbard, T.J.P. et al. “Ensembl 2007.” *Nucleic Acids Research*, 2006, 0: D1-8.
11. Enright, A.J. et al. “MicroRNA targets in *Drosophila*.” *Genome Biology,* 2003, 5: R1.
12. Betel, D. et al. “The microRNA.org resource: targets and expression.” *Nucleic Acids Research*, 2008, 36: D149-D153.
13. John, B. e al. “Human MicroRNA Targets.” *PLoS Biology*, 2004, 2(11): e363.
14. Ioshikhes, I. et al. “Algorithms for mapping of mRNA targets for microRNA.” *DNA and Cell Biology*, 2007, 26(4): 265-272.
15. Lewis, B.P. et al. “Prediction of mammalian microRNA targets.” *Cell*, 2003, 115: 787-798.
16. Lewis, B.P. et al. “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.” *Cell*, 2005, 120: 15-20.
17. Krek, A. et al. “Combinatorial microRNA target predictions.” *Nature Genetics*, 2005, 37(5): 495-500.
18. Saetrom, O. et al. “Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms.” *RNA*, 2005, 11: 995-1003.
19. Koza, J.R. *Genetic programming: on the programming of computers by natural selection.* MIT Press, Cambridge, MA. 1992.
20. Meir, R. and Ratsch, G. “An introduction to boosting and leveraging.” In *Advanced lectures on machine learning* (eds. S. Mendelson and A. Smola), Vol 2600, pp. 118-183. Springer-Verlag, GmbH.
21. Kertesz, M. et al. “The role of site accessibility in microRNA target recognition.” *Nature Genetics*, 2007, 39(10): 1278-1284.
22. Lytle, J.R. et al. “Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5’ UTR as in the 3’ UTR.” *PNAS*, 2007, 104: 9667-9672.

23. Duursma, A.M., et al. “miR-148 targets human DNMT3b protein coding region.” *RNA*, 2008, 14: 872-877.

24. Liu, G. et al. “Pre-miRNA loop nucleotides control the distinct activities of miR-181a-1 and mir-181c in early T cell development.” *PLoS One,* 2008, 3(10): e3592.