A Comprehensive Study of Target Prediction Algorithms for Animal MicroRNAs(miRNAs)

Joyshree Nath
A.K.Chaudhuri School of IT,
Calcutta University, Kolkata, India

Asoke Nath
Department of Computer Science
St. Xavier's College(Autonomous), Kolkata, India

ABSTRACT

The discovery of microRNAs has been a path-breaking step in understanding the full scope of post-transcriptional gene regulation. The microRNAs (miRNAs) are highly conserved, non-coding short ribonucleic acid (RNA) molecules, approximately 22 nucleotides long[6] and are found in all eukaryotic cells, except fungi and marine plants. MicroRNAs (miRNAs) post-transcriptionally regulate the expression of target genes by binding to complementary sequences on target messenger RNA transcripts, usually resulting in translational repression and thus inhibition of the expression of target mRNAs. Complete complementarity between miRNA:mRNA pairs is rare in mammals, but as little as a 6 bp match with the target mRNA can be sufficient to suppress the gene expression[8]. MicroRNAs, which were initially determined as moderate biological modifiers, have now emerged as powerful regulators of diverse cellular processes with important roles in tissue remodeling[9]. It throws light into the causes of diseases like lymphoma, leukemia, cancers and many cardiac problems where miRNA:mRNA pairing is found to play crucial roles[5]. Many computational methods are being developed to identify the relationship between the animal miRNAs and their target mRNAs. Here we study two of those recent methods to identify the target mRNAs of existing animal miRNAs.

Keywords: miRNA, remodeling, nucleotides, mRNA

1. INTRODUCTION

MiRNAs are a class of small, non-coding regulatory RNAs that are important in post-transcriptional gene silencing[7]. They facilitate deadenylation, which leads to rapid mRNA decay[11]. 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 differentiation, cell death, hematopoiesis, and oncogenesis.There are many computational algorithms developed to predict plant and animal miRNA targets. Since in plants we find almost perfect sequence complementarities, researchers used near-perfect complementarities to predict plant miRNA targets. But regarding animals very few miRNA targets could be predicted, because of limited sequence complementarities between the animal miRNAs and their gene targets[7]. Most of the different computational programs for animal miRNA target prediction, including the very popular miRanda(Enright et al. 2003, John et al. 2004), TargetScan(Lewis et al. 2003), and PicTar(Krek et al. 2005), mainly rely on the identification of the seed region between the miRNA and the corresponding target genes. But, the presence of a seed region, although conserved across evolution, is not a reliable way to identify functional miRNA targets. A significant proportion of the predicted miRNA–mRNA target pairs present in appropriate seed regions, are false positives (Lewis et al. 2005; Didiano and Hobert 2006)[4]. Again, since each miRNA can target several mRNA sequences and the target sites may overlap to some degree[2], it is a very complex and laborious task to identify these targets correctly. So more different types of computational target prediction tools are needed to build models with high specificity to accurately predict miRNA targets. Here we study some of the recent different approaches for prediction of targets of animal miRNAs and their results.

2. BACKROUND STUDY

2.1. Biogenesis

MiRNAs are a class of short non-coding RNAs. MiRNAs originate in the nucleus and modify mRNA in the cytoplasm. Each miRNA is first transcribed and long primary transcripts (pri-miRNAs) are generated. The pri-miRNA is processed in the nucleus itself into hairpin precursors of 60 to 70 nt by the RNase III–type enzyme Drosha and an RNA binding protein Pasha. This stem-loop precursor molecule, known as pre-miRNA, is transported to the cytoplasm. by the RNase III protein Dicer. In the cytoplasm, they are processed into unstable, 20 to 25 nt miRNA duplex structures by the RNase III protein Dicer. The pre-miRNA attaches to the multiple-protein nuclease complex RISC (RNAi-induced silencing complex). This complex degrades one of the strands, passenger strand, leaving the other RNA strand, the mature strand, to bind to its target mRNA. After the mature miRNA binds to mRNA, RISC is freed to find and process another pre-miRNA. The target sequence typically resides in the 3'UTR region of the mRNA, although some have been found in the 5'UTRs and in coding regions. The function of a miRNA is ultimately defined by the genes it targets and the effects it has on their expression[10],[5].

3. STUDY OF SOME IMPOTANT MIRNA TARGET PREDICTION ALGORITHM: DEPICTING THE INDIVIDUAL METHODS

To build a correct and appropriate computational method for identification of animal miRNA target genes, is a complex job. It is due to the fact that animal miRNAs display limited sequence complementarity to their gene targets. It is difficult to build a fully efficient target prediction model without any loopholes. Nevertheless there are some effective target prediction algorithms for mammalian target identification. Now we will be discussing some of those algorithms:

3.1. NbmiRTar :

Around 2007,Malik Yousef, Segun Jung, Andrew V. Kossenkov, Louise C. Showe and Michael K. Showe[3] presented a machine-learning approach for predicting miRNA target sites based on the Naïve Bayes (NB) classifier. This method was called NBmiRTar and unlike many other methods it did not need sequence conservation. Here along with the seed, the out-seed segments of the miRNA:mRNA duplex were also used for target identification. A training dataset was prepared by the authors to experiment with and get proper results.

3.1.1. Selection of the Training dataset and features

The training dataset contained 225 confirmed miRNA targets (human, mouse, fruit fly worm and zebrafish) and 38 confirmed false target predictions to serve as positive and negative examples, respectively. Some additional negative examples were generated by the miRanda algorithm to make the algorithm function efficiently. Few miRNA features were chosen to device this machine learning approach.

The features chosen in this study were based on the following assumptions:

(1) The complementarity of 7–8 bases in the seed region(5′ 8 nt of the miRNA) are sufficient for proper miRNA:mRNA duplex formation.

(2) A seed segment with weak complementarity can be compensated for by the out-seed(3′ remainder portion of the miRNA) sequence to make a functional duplex.

(3) Good complementarity in the out-seed region alone is not sufficient for functional duplex formation.

For each part of the duplex (the seed and the out-seed) the following features were considered to form 57 structural features to help in target identification:

(1) The number of paired bases, (2) The number of bulges, (3) The number of loops, (4) The number of asymmetric loops, (5) Eight features, each representing the number of bulges of lengths 1–7 and those with lengths greater than 7, (6) Eight features, each representing the number of symmetric loops with lengths 1–7 and those with lengths >7, (7) Eight features each representing the number of asymmetric loops with lengths 1–7 and those with lengths >7 and (8) The distance from the start of the seed (the 3′ end) to the first paired base of the 5′ start of the out-seed part.

3.1.2. The Procedure

The **miRanda program**(John et al. 2004) assigns a score to each pairwise alignment to describe the maximal local complementarities. A score of +5 for G:C and A:T pairs and +2 for G:U wobble pairs is assigned over here. The final miRanda score is computed as the sum of all these single-residue-pair match scores over the entire duplex structure. The output of miranda comes to the **NB(Naïve Bayes) classifier**. It calculates the probability of a given example belonging to a particular class, assuming that the features constituting the example are conditionally independent given the class. Although the NB classifier had high accuracy at finding known miRNA:mRNA target genes, the number of false positives were not much reduced. So, a proper threshold value (0.9 as taken by the authors) was applied to further reduce the number of false positive predictions. The classifier assigned a score to each miRNA: mRNA duplex and classified it into the positive class (target) and the negative class (non-target).

3.2. Mtar:

A method called MTar(Chandra et al,2010)[5] was proposed for human transcriptome**.** This model considers evolutionary conservation analysis, incorporates 16 very important positional, thermodynamic and structural features (identified from the wet lab proven miRNA:mRNA pairs) for miRNA target identification, and classifies the training dataset into three miRNA target classes (5' seed-only, 5' dominant, and 3' canonical).

3.2.1. Selection of the Training dataset and features

Only experimentally verified microRNAs by wet lab and their targets were considered for this method. The targets whose exact binding site were not verified were excluded from the dataset to maintain the quality of the data used. The collected dataset consisted of 882 human records for 741 genes by 138 miRNAs. The training dataset was classified into three portions based on the three target classes (5' seed-only, 5' dominant, and 3' canonical).

The **5' dominant seed site targets** (5' seed-only), possessing high complementarities in 5' end and a few complementary pairs in 3' end. 2) The **5' dominant canonical seed site targets** (5' dominant), possessing high complementarities in 5' end (of the miRNA) and very few or no complementary pairs in 3' end. 3) The **3' complementary seed site targets** (3' canonical) have high complementarities in 3' end and insufficient pairings in 5' end.

The final training dataset contained 40 positives and 56 negatives for 3' canonical target class, 58 positives and 74 negatives for 5' dominant target class and 52 positives and 70 negatives for 5' seed-only target class.

The authors analysed and considered 16 parameter features to be very relevant to their approach. These miRNA target site features, were divided into three categories-(a) **structural** (Seed score, Out seed score, WC pairs, Wobble pairs, Mismatches, Length-bulge,Number-bulges, Proportion), (b) **thermodynamic** (Free energy, Hybridization Energy, Normalized free energy, Difference in hybridization energy) and (c) **positional** (Positional pair score,Matrix score,Deviation matrix score,Deviation positional score).

3.2.2. The Procedure

1) Firstly, the miRNA sequence input were aligned with the mRNA target sequence using a modified Smith-Waterman local alignment algorithm, which preferred mismatches to gaps and assigned higher penalty for gaps. The scoring scheme was to assign each Watson-Crick(WC) pair a score of 5, each G:U pair , a score of 1 and all others a score of -3.

2) Secondly, miRNA:mRNA duplex was checked for seed and out-seed complementarity. The complementarity score in the seed region and out seed region were calculated to classify the target candidates into 3 types. These classes that were considered to be processed by the method were:

3) (a) 5' seed-only: Here minimum 6 WC pairs and no wobble or mismatch were allowed in the seed region.

The non-seed region may contain a minimum of 4 matching pairs including G:U pairs; (b) 5' dominant: Here minimum 5 WC pairs, with one mismatch and a maximum of 2 G:U pairs were allowed in the seed region. Minimum 5 matching pairs including G:U pairs should be in the non-seed region and (c)3' Canonical: Here minimum 3 WC pairs, 4 mismatches and maximum 3 G:U pairs were allowed in the seed region. The non-seed region should contain a minimum 7 matching pairs including G:U wobble pairs.

4) The potential target candidates for the miRNA belonging to the three different categories (5' seed-only, 5' dominant and 3' canonical) were located by aligning each segment with the miRNA.

5) Then, appropriate weights were assigned to those 16 most relevant parameters for each candidate.

6) This output goes to an Artificial neural networks (ANN) classifier for target validation. This classifier was used to verify the miRNA targets, due to their ability to deal with complex non-linear data. Three separate ANNs for each target class (5' seed-only, 5' dominant and 3' canonical) were trained to validate the target candidates of each of those classes.

4. RESULTS OBTAINED FROM EACH METHOD:

4.1 NBMiRTar[3]

I. This method could yield a specificity of 0.99 and a sensitivity of 0.94 using 900 negative examples.

II. In this method the authors compared the number of target predictions generated by miRanda(John et al. 2004), the NB classifier. For 10 known human miRNAs, the NBmiRTar tool had a reduction of 75% of miRanda predictions with a recovery rate of 77% of the confirmed targets. Finally as an output this method gave 620 757 predictions from the 10 known human miRNAs.

III. For a single human miRNA, mir-15, miRanda produced 88 376 predictions which was subsequently reduced to 3479 predictions after applying this method.

4.2 MTar[4]

I. MTar was tested with all the three parameters combined and also with different pairwise combinations of these features. Mtar predicted a total of 2663 target sites including 819 experimentally verified targets of 129 miRNAs.

II. The authors compared the results obtained from MTar with few other methods (MiRanda, TargetScan, RNA22, PicTar, MiTarget). In their experiment MiRanda gave a specificity of 82%,PicTar (~ 70%) and TargetScan (~ 80%) by the same test dataset used for MTar. MTar had an average accuracy of 92.8%, sensitivity of 94.5% and a specificity of 90.5% for the miRNA targets for all 138 experimentally verified miRNAs in human genome.

5. MERITS AND DEMERITS OF THE METHODS DISCUSSED:

5.1 NBMiRTar[3]

i. NBmiRTar maintained the levels of sensitivity indicating the importance of the extracted features.

ii. Many target prediction tools either predicted very large numbers of miRNA targets making biological validation very difficult or produced smaller numbers of predictions but only with highly conserved sequences. NBmiRTar does not rely on conservation and significantly minimizes the number of target candidates to be tested.

iii. NBmiRTar demonstrates both high specificity and high sensitivity and thus a high accuracy in target identification.

iv. This method uses a threshold value(0.9) for reducing the false positive predictions hugely but loses some of the true miRNA targets in this procedure.

v. The number of target predictions is considerably reduced while retaining the sensitivity of the procedure.

5.2 MTar[4]

i. The performance of MTar was better in accuracy than MiRanda, TargetScan, RNA22, PicTar or MiTarget which are few of the popular and existing algorithms of today.

ii. This method, unlike others, identifies the three types of targets (5' seed-only, 5' dominant, and 3' canonical) in a single framework.

iii. Target site multiplicity and cooperativity are handled very effectively.

iv. Target identification is based on the selection of the 3 different features (positional, thermodynamic and structural).

v. MTar still gave false positives and lowering it further effected on its sensitivity.

vi. This method is a complex procedure in comparison to few other existing algorithms.

6. CONCLUSION AND FUTURE SCOPE

NBmiRTar is a machine-learning approach to miRNA target prediction that does not rely on sequence conservation and is still able to significantly reduce the number of target predictions while retaining an acceptable sensitivity. MTar can identify all known three types of miRNA targets (5' seed-only, 5' dominant, and 3' canonical). The performance of MTar was compared against existing solutions and the method is found to be more accurate.

Since the first lin-4 discovered miRNA lin-4 (Lee et al. 1993), numerous approaches contributed greatly to understanding these microRNAs. The methods we have discussed are very recent approaches. To give the ultimatum that whether any of these is totally effective or totally useless will be too early to comment. There are methods that provide prediction by using

multiple algorithms. But many such combinatorial predictions perform worse than the prediction by one accurate algorithm, because of the trade-off between specificity and sensitivity (Alexiou et al., 2009). Since existing target prediction algorithms rely on different assumptions and approaches, one must check the underlying assumptions and limitations first before employing a target prediction tool. So one can choose NBMiRTar or MTar according to one's requirements.

There is continuous thrive to get an even more efficient computational tools for a more effective and accurate target prediction . More and more path-breaking biological insights will lead to the creation of new algorithms based on mechanistic understanding. Although knowledge of miRNAs has accumulated rapidly in recent years, still many stones are left unturned and much of the miRNA functions in the biological network needs to be discovered.

7. REFERENCES

[1] MicroRNA targets in *Drosophila,* Anton J Enright**,** Bino John**,** Ulrike Gaul**,** Thomas Tuschl**,** Chris Sander and Debora S Marks, *Genome Biology* 2003, 5:R1**.**

[2] Human MicroRNA Targets, Bino John, Anton J. Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, Debora S. Marks, PLoS Biol 2(11): e363(2004).

[3] Naïve Bayes for microRNA target predictions—machine learning for microRNA targets, Malik Yousef, Segun Jung, Andrew V. Kossenkov, Louise C. Showe and Michael K. Showe,Volume 23, Issue22,Page. 2987-2992(2007).

[4] MicroRNA target prediction by expression analysis of host genes, Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avellino, Nicola Meola, VincenzaMaselli, SantoshAnand, Luisa Cutillo, Andrea Ballabio, and SandroBanfi, Genome Res, v.19(3); Mar,(2009).

[5] MTar: a computational microRNA target prediction architecture for human transcriptome, Vinod Chandra**,** ReshmiGirijadevi, Achuthsankar S Nair**,** Sreenadhan S Pillaiand Radhakrishna M Pillai, BMC Bioinformatics **11**(Suppl 1):S2(2010).

[6] Computational miRNA Target Prediction in Animals by Leyan Tang.

[7] Got target?: computational methods for microRNA target prediction and their extension,Hyeyoung Min and Sungroh Yoon, Exp Mol Med. April 30; 42(4): 233–244(2010).

[8] Experimental strategies for microRNA target identification, Available at: nar.oxfordjou**rna**ls.org/content/early/2011/06/07/nar.gkr330.full

[9] The Art of MicroRNA Research, available: http://circres.ahajournals.org/content/108/2/219.abstract

[10] miRNA Biogenesis: http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/epigenetics-noncoding-rna-research/Epigenetics-Learning-Center/miRNA/miRNA-Biogenesis.html

[11] MicroRNAs direct rapid deadenylation of mRNA: www.pnas.org/content/103/11/4034.full