Fall 2010

Computational miRNA Target Prediction in Animals

Leyan Tang
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the [Other Computer Sciences Commons](#)

Recommended Citation

Tang, Leyan, "Computational miRNA Target Prediction in Animals" (2010). *Master's Projects*. 14.
DOI: <https://doi.org/10.31979/etd.dbfv-wku2>
https://scholarworks.sjsu.edu/etd_projects/14

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Computational miRNA Target Prediction in Animals

A Thesis

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Computer Science

by

Leyan Tang

Dec 2010

This dissertation has been approved
for the Department of Computer Science
San Jose State University

Dr. Sami Khuri, Department of Computer Science, SJSU

Dr. Robert Chun, Department of Computer Science, SJSU

Dr. Xiudian Fang, Department of Computer Science, UVA

ACKNOWLEDGMENT

I would like to thank Dr. Sami Khuri, not only for his guidance, encouragement and patience though out every class I have taken with him including this writing project, but also introducing me to the fascinating field of Bioinformatics and teaching me essentials of Hidden Markov Models. My gratitude to Dr. Robert Chun and Dr. Xiuduan Fang for serving on my defense committee.

Thank you to my family and friends, Yang Liu, Yue Chen, Lin Huang and Yang Li for the support. Special thanks to James O. Bowlby, who provided both hardware and software support to this project and spent long hours providing technical support.

ABSTRACT

by Leyan Tang

miRNAs are a class of small RNA molecules about 22 nucleotides long that regulate gene expression at the post-transcriptional level. The discovery of the second miRNA 10 years ago was as much a surprise in its own way as the very structure of DNA discovered a half century earlier[1]. How could these small molecules regulate so many genes? During the past decade the complex cascade of regulation has been investigated and reported in detail[2]. The regions of the genome called untranslated regions, or UTRs, proved true to their name: they were indeed untranslated, but certainly not unimportant: they act as the origin and often the destination of miRNAs.

miRBase[3] contains 1048 human miRNAs with more undoubtedly on the way. But experimental identification of miRNA targets has proven dreadfully slow and difficult. Instead, scientists have turned to computational target prediction programs as the preferred method to quickly identify potential miRNA targets. Current prediction tools have produced a huge number of potential target sites, but determining if they are correct, or which algorithms produce the most reliable predictions, remains an open question.

This project examines one type of algorithm, a probabilistic model called a profile Hidden Markov Model (pHMM), and uses it to predict miRNA target sites. HMMs are known to be very effective in pattern recognition and have been successfully applied to various bioinformatic applications, such as gene finding, multiple sequence alignment and protein family classification[4]. We proposed to build a pHMM from known miRNA interactions and use this model to identify potential miRNA target sites in UTR regions by abstracting the Watson-Crick base pairs into meta codes intended to more naturally describe important relationships in RNA folding. High quality positive training data came from the best curated mRNA:miRNA data-bases we could find, while negative training data was generated using random sequences. The purpose of this project was to demonstrate the flexibility of the pHMM architecture to process many kinds of interesting data and by doing so improve their miRNA target site prediction.

Table of Contents

I. INTRODUCTION	1
1.1 Background.....	2
1.2 Biogenesis.....	2
1.3 Mechanism	4
1.4 Function.....	6
1.5 Overview of existing animal miRNAs prediction tools.....	7
1.6 Overview of existing miRNA target prediction tools	8
1.7 Future Challenges	10
II. PRELIMINARY WORK	12
2.1 1st Order Markov Chain.....	13
2.2 Data	17
2.3 Rank the computational predicted miRNA targets	18
III. Human miRNA targets PREDICTION.....	20
3.1 Blast Scheme:	21
3.2 Seed Matching Scheme	32
3.3 Hidden Markov Model.....	34
III. CONCLUSION	38
VI. FUTURE WORK	39
REFERENCES.....	40
APPENDICES	43

List of Tables

Table 1: Rules for Converting the Binding Sequence	13
Table 2: PWMs for Groups of miRNA Length 18 – 25 nt	14
Table 3: Emission Probabilities for the 22 State 1st Order Markov Model	15
Table 4: Multiple Binding Sequences Alignment	17
Table 5: FASTA Style of Various Databases	30
Table 6: Candidate Targets Sites for a Single 7-mer Seed Matching	33
Table 7: One profile calculated for a set of positive training data	35
Table 8: Positive data V.S Negative data	36
Table 9: Four-Fold Cross Validation Result	37

List of Figures

Figure 1: Schematic Overview of miRNA Biogenesis[8]	3
Figure 2: Two Typical Classes of Animal miRNA Target Sites [49]	5
Figure 3: Topologies of (a) the 1st order Markov Model and (b, c) hidden Markov Models	16
Figure 4: Project Work Flow	21
Figure 5: Blast Scheme Work Flow	22
Figure 6: CPU Usage Before Parallel Process	27
Figure 7: CPU Usage During Parallel Processing.....	28
Figure 8: Sample Blastn Output File from PACdb	29
Figure 9: Feasibility Test Diagram.....	31
Figure 10: Single miRNA Seed Site has 35 Possible Target Sites	33
Figure 11: The pHMM Topology	34

I. INTRODUCTION

This project predicts human miRNA targets in two steps. The first step scans databases looking for potential miRNA target sites according to given miRNA sequences, and the second step employs profile HMMs to distinguish those target sites most likely to be true. The second part can also be used to assess diverse prediction results from different tools. miRNAs belong to a class of non-coding RNAs about 22 nucleotides long which bind to target mRNAs in order to down regulate gene expression at the post-transcription stage[5]. Studies have shown that miRNAs play a significant role in both cancer and viral infections. Finding miRNA target sites is expected to be fundamental for novel anticancer and viral therapies[6]. Since identifying miRNAs in wet lab is dreadfully difficult and time consuming, computational target prediction programs remain the most efficient way of finding potential miRNA target sites on mRNAs[7]. Though current prediction tools produce a huge number of potential target sites, to determine which algorithm produces the most reliable prediction results remains an open problem.

This project examines one type of algorithm, a probabilistic model called a profile Hidden Markov Model (pHMM), and uses it to assess miRNA target sites. HMMs are known to be very effective in pattern recognition and have been successfully applied to various bioinformatics applications, such as gene finding, multiple sequence alignment and protein family classification. We propose to build a pHMM from known miRNA interactions and use this model to identify potential miRNA target sites in UTR regions by abstracting the Watson-Crick base pairs into meta codes intended to more naturally describe important relationships in miRNA to mRNA binding. The positive training data came from tarbase[8] and miRNAMap[9] which claim their data are experimentally verified, while negative training data

was generated from the same data set with random mutations. The purpose of this project was to enhance current human miRNA prediction tools by using a pHMM architecture. It also demonstrated the flexibility of the pHMM architecture to process many kinds of interesting data.

1.1 Background

miRNAs are a class of short non-coding RNAs first discovered in 1993 and intensely investigated a decade ago. They are post-transcriptional genetic regulators that play a crucial role in eukaryotic gene expression and regulation. The function of a miRNA is ultimately defined by the genes it targets and the effects it has on their expression. Unlike the situation for plants, animal miRNA targets are difficult to predict because there exist some mismatches, gaps and G:U wobble pairs in miRNA:mRNA duplexes. Furthermore, some animal miRNAs have multiple targets on the same mRNA[7].

This section will present a brief description of miRNA biogenesis, and then an overview of computational tools that predict the miRNA itself and its target. It will also present the context out of which this project grew. Reading papers written over the last ten years one can clearly see the intense, intelligent, and creative debate over how to best couple computational and laboratory methods. It is hoped that this project can contribute to that conversation.

1.2 Biogenesis

Figure 1: Schematic Overview of miRNA Biogenesis[8]

miRNAs originate in the nucleus and modify mRNA in the cytoplasm. Figure 1 gives a good summary of their biogenesis in mammals. They reside in the introns of genomic DNA, i.e., the portions of a gene normally excised by spliceosome. Each miRNA is coded as approximately 300 bases in length and may either contain its own promoter region (in the case of antisense miRNA) or become parasitically transcribed when the gene in which it is embedded is itself transcribed. At this stage, the miRNA is called a pri-miRNA and acts like any other transcript: it is capped with a modified guanine nucleotide at the 5' end and polyadenylated at the 3' end. Due to internal pairing, two strands of the pri-miRNA hybridize forming a single-stranded loop, a double-stranded stem, and two single-stranded tails.

The hybridized stem is recognized by a group of proteins called the Microprocessor Complex. Two of the largest proteins are shown in Figure 1: DGCR8 and Drosha. This complex

identifies the end of the stem and cleaves the two dangling single-stranded nucleotides, resulting in the much smaller 60 – 70 base precursor molecule called the pre-miRNA. This precursor is transported out of the nucleus by proteins Exportin-5 and Ran-GTP. In the cytoplasm, a complex called TRBP binds to the single-stranded loop structure and Dicer mediates its cleavage. The result is a much shorter, approximately 22 base double-stranded RNA fragment.

Not shown is an alternative pathway called mirtrons produced directly from the nuclear splicing machinery. They are not processed by Drosha but continue further processing like other pre-miRNAs.

1.3 Mechanism

The pre-miRNA attaches to a protein complex called RISC (RNA-induced silencing complex) containing argonaute proteins such as Ago2. This complex degrades one of the strands, called the passenger strand, leaving the other RNA strand, called the mature strand, to bind to its target mRNA. The mature strand stays attached to RISC until it actually finds complementary regions on mRNA. After the mature strand 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 even in coding regions. Binding generally occurs in organelles called P-bodies before mRNA arrives at ribosomes for translation. P-bodies contain an enzyme that degrades any mRNA perfectly complementary to miRNA. If the match is not perfectly complementary, i.e. if some bases differ or a bulge forms on the mature strand, then the mRNA is left intact and released from the P-body. When it eventually reaches the ribosome it

will not be translated into a protein because the miRNA quite effectively blocks ribosomal processivity. Although the function of miRNA is generally to down regulate mRNA, and thus gene expression, by mechanisms not yet understood some miRNAs can upregulate genes as well[15].

Figure 2: Two Typical Classes of Animal miRNA Target Sites [49]

Whereas most plant miRNAs completely hybridize with their target, mammalian miRNAs generally evolve mismatches or buldges. Figure 2 shows two typical classes of animal miRNA target sites: either perfect binding in the seed region and mismatches at the 5' end; or a

single bulge in the seed region with compensatory perfect binding at the 5' end. The majority of animal miRNA target sites belong to the first class which has a perfect binding seed site and several mismatches and bulges.

This creates significant problems when identifying putative miRNAs using strictly computational methods. Only a few bases match identically, and searching a large genome (e.g.: three billion bases) for a few matches will generate many hits simply by chance. To improve accuracy, current computational methods utilize sequences that are highly conserved between species. However, this skews algorithmic discovery in the direction of previously identified highly conserved sequences [14, 15, 16]. Fortunately, several viral and human nonconserved miRNAs have recently been experimentally verified [17, 18]. The question remains, can algorithmic methods alone discover both conserved and nonconserved miRNAs?

1.4 Function

miRNAs play an important role in gene expression by regulating a large number of target genes. Currently, only a limited number of the hundreds of animal miRNA genes have been genetically identified and linked to definite functions. The majority of miRNAs' target genes are computationally predicted and still need to be verified. Each mammalian miRNA may regulate approximately 200 target genes[43]. Indeed, some of the earliest discovered miRNA genes in *C.elegans* were *let-7* and *lin4*, now confirmed to manage developmental timing, and *lisy-6* which regulates the left-and-right asymmetry in the nervous system [44]. For human and other vertebrate cell lines, miRNA genes are involved in tumor suppression, antiviral defense, adipocyte differentiation and susceptibility to cytotoxic T-cells[43].

1.5 Overview of existing animal miRNAs prediction tools

Once the ubiquity of miRNAs was fully understood, intense activity ensued to identify the miRNAs themselves as well as their mRNA targets. Initially, it appeared to be a straight forward annotation problem summarized by the following rules [20]:

1. Experiments should confirm expression of a ~22 nucleotide transcript.
2. Folding calculations should not find large bulges, or
the fold should be conserved within the phylum, or
Dicer impaired organisms should accumulate more of the transcript.

It soon became clear that these rules were too simple to unambiguously classify miRNA from other functional or simply degraded transcripts.

Computational activity shifted to investigating the unique structure of the stem since it necessarily formed Watson-Crick pairs with itself within a very short distance. It was quickly noticed that the stem structured appeared highly conserved across many organisms, so orthologous structures were identified in high number [21]. Orthologous structures continue to dominate search strategies for both miRNA and targets, but many lament that these strategies unfairly skew the reported results [22].

The stem structures need not match perfectly. Therefore, simple text searching or pattern matching is confounded by the multiple insertions or mismatches encountered in real miRNAs. More complicated classifiers were attempted, such as thermodynamic folding calculations (e.g.: MirScan[21] and PalGrade [22]), HMMs (described below) and support vector machines (SVMs). SVMs are binary classifiers that, unlike HMMs, do not work with probabilities. Instead they embed training data in a large number of dimensions and identify a hyperplane that

provides the best separation. In some cases, they can tease out hidden structure and have been used for feature prediction [23] and Drosha interaction prediction [24].

1.6 Overview of existing miRNA target prediction tools

The flip side of computational miRNA prediction is an automated method of identifying the target sites on mRNA. In many ways, this is viewed as a more constrained problem because one can usually restrict the search for targets to the 3' UTR region of known fully formed mRNA. In addition, there are often multiple target sites in this region. However, the search has proven challenging because animal miRNA:mRNA duplexes often contain mismatches, gaps, bulges or G:U wobble bases. These structures confound more traditional high speed search algorithms such as BLAST. More complicated local alignment algorithms such as Smith-Waterman are computationally intense and therefore unusable for searching large databases.

In order to use computationally intensive algorithms, the search space needed to be reduced. One method constructed a database containing only 3' UTRs conserved between two species of *Drosophila*[27]. It was found that the 8 nucleotides at the 5' end showed good complementarity to miRNA with few mismatches or wobble pairs, and several target sites were predicted.

MiRanda [28] extended this technique to all miRNAs in *Drosophila*, and it was later applied to humans as well [28], indicating early on that ortholog searches were a powerful tool for discovering miRNAs. TargetScan [30] pioneered the idea of finding seeds, i.e.: 7 nucleotide segments in the 3' UTR regions of many organisms that should be perfectly complementary to positions 2 – 8 in miRNA. Once a seed was found it was extended allowing for mismatches,

bulges, or wobble bases. The same team continued their work by adding more organisms to the working set and reduced the seed count to 6 nucleotides.

Known miRNA targets in *C. elegans* were searched for high GC content [31], and consecutive high scoring regions were termed a “binding nucleus”. By adjusting Watson-Crick and wobble pairing weights, the differences between scores of targets and randomized non-targets could be optimized. This classifier was then applied to de novo sequences.

DIANA-microT [32] predicted targets in humans. Ten conserved miRNAs were found in both mouse and human. Again, the first few nucleotides at the 5’ end proved to be the only bases with full complementarity.

A dynamic programming algorithm was developed called RNA-hybrid [33]. It allowed the user to specify a seed and computed the lowest free energy between the miRNA and mRNA strands. The method was also applied to orthologs in related species.

PicTar [34] starts by finding almost perfect 7 nucleotide seeds starting at the 5’ end of the miRNA, filters them by free energy, and calls the results “anchors”. Since mRNA usually contains multiple targets, an HMM then selects only multiple target genes.

MovingTargets [35] builds a database of possible targets by finding orthologs of 50 nucleotides or less in the 3’-UTR of closely related species. Users could change any of five criteria, including: how many target sites, free energy, consecutive nucleotides from the miRNA’s 5’ end and number of wobble pairs.

A Bayesian network attempted to classify conserved sites from evolving sites [36]. Starting with seeds it found the probability that a target is still evolving compared to a set of conserved orthologs.

TargetBoost [37] uses more contemporary combination of boosting and machine learning to uncover the hidden rules of target hybridization. Several general patterns, themselves a classifier, are “spawned” and then evolve programmatically. The classifiers are combined with weights that were “boosted” depending on how well they performed.

If miRNA always down regulates or destroys mRNA then putative targets could be correlated with reduced gene expression. Several experiments attempted to capitalize on this approach, but eventually it was discovered that the relationship between mRNA expression and the number of targets was not straightforward [38].

The mRNA secondary structure of the target was also investigated, and eventually a method called PITA computed the change in free energy between the unbound and bound target. [39]

1.7 Future Challenges

The need to identify more and more accurate miRNA targets will likely continue for some time, albeit in the light of continual new information which may make the process even more challenging. For example, a) a process called RNA editing changes the miRNA after it is transcribed, so the relationship between miRNA and mRNA will be more difficult to establish [40]; b) additional studies regarding the base composition of the 3'-UTR region are needed to understand why targets seem to prefer more AT base pairs than GC [41]; and c) more investigation is needed to clarify the specific role of promoters, Drosha and Dicer processing, and evolutionary models of the 3'-UTR region and the pre-miRNAs [42].

Although an incredible amount has been learned since miRNAs were first discovered, a comprehensive and accurate enumeration of all miRNAs along with their targets is still a goal for

the future. Until that day, new data will become available and new algorithms will be evolved. It is hoped that the algorithm proposed in this project will be useful and make a contribution to this evolution.

II. PRELIMINARY WORK

The hypotheses is miRNA:mRNA interactions have discoverable binding patterns. One paper claims that a 1st order Markov model is capable of assessing potential miRNA targets by converting each miRNA:mRNA interaction into a 0-1 sequence. Looking at the miRNA side only, Watson-Crick base pairings count as 1, the rest of the pairings are all 0s. The limitation of this 0-1 binding pattern sequence is it is unable to distinguish different miRNA:mRNA interaction structures. As shown in Table 1, (a) has an interior loop at the center and 22 nucleotides in its miRNA sequence; (b) has perfect binding with no interior loop and the miRNA sequence is only 14 nucleotides long; (c) a 22 nt miRNA binds to a 14 nt target site; and (d) a 14 nt miRNA binds to a 22 nt target sites. However, as different as these four cases are, all of them share the same 0-1 binding pattern sequence[45].

In order to make the pattern more specific, we changed the coding rules. Starting from the miRNA side, there are four types of base-pairings to the target site on mRNA: a Watson-Crick pair, a G-U wobble pair, no pair, and a bulge. In order to better examine the binding patterns, every pair of aligned miRNA:mRNA sequences is converted to a new binding pattern as shown in Table 1, where ‘N’ is for no pair, ‘P’ is for Watson-Crick pair, ‘W’ is for G-U wobble pair, and ‘B’ is for a nucleotide in the miRNA strand with a gap in mRNA strand. In this way, each miRNA interaction can be converted into a unique binding pattern sequence, similar to techniques used to model protein secondary structure. Based on these binding pattern sequences, we built our own series of Markov Models to examine more realistic miRNA:mRNA interactions.

a.	target	5'
A	G	C U C A G A G G U C U 3'
:		: : : :
U U G A		G G A G A U G G A G
miRNA	3'	U A U G U U U 5'
Pattern	P W P P N N N N N N	P P P N P P P P P P P P
b.	target	5' U 3'
A G C U		C C U C U A C C U C
:		: : : :
U U G A		G G A G A U G G A G
miRNA	3'	U 5'
Pattern	P W P P P P P P P P P P	
c.	target	5' U 3'
U U G A		G G A G A U G G A G
:		: : : :
A G C U		C C U C U A C C U C
miRNA	3'	U A U G U U U 5'
Pattern	P W P P B B B B B B P P P B P P P P P P	
d.	target	5' A A G C U C A G A G G U C C U C C U A C C U C U 3'
:		: : : :
U U G A		G G A U G A U G G A G
miRNA	3'	
Pattern	P W P P P P P N P P P P P P P	
0-1 Pattern	1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1	

We analyzed common characteristics of binding patterns by using position weight matrices (PWM) and 1st-order Markov Model. PWMs were built for known human miRNA genes grouped by length as shown in Table 3. The test data came from miRNAmap[13], which claims their data comes from experimentally verified human miRNA interactions. It is a well-known fact that miRNA nucleotides 2 to 7 bind perfectly with their target mRNA to form the "seed" or "nucleus". Beyond this region, G:C and A:U paired nucleotides constitute around

50% when compared to other types of binding such as mismatches or bulges. While the PWM faithfully yields the types of base-pairings, it does not give the variations of bindings as we move from one nucleotide to the next. Therefore, we modeled, studied and analyzed this variation using a 1st-order Markov Model.

TABLE 2: PWMs FOR GROUPS OF miRNA LENGTH 18 – 25 NT

22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'	
48	48	48	80	81	108	140	152	122	140	144	210	280	278	3	0	0	0	0	0	1	2	4	5
472	482	303	374	392	281	229	307	381	388	318	288	220	308	4	0	19	2	8	0	3	338	W	
779	473	407	384	808	848	812	488	808	487	482	818	838	888	0	9	4	3	0	0	4	2	1088	N
838	1131	1078	1314	1183	1200	1281	1208	1144	1137	1181	1138	1017	888	2128	2123	2109	2127	2127	2127	2128	734	F	

23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'
22	20	19	27	40	53	59	88	83	73	84	115	139	110	0	0	0	0	0	0	1	4	5	
200	200	208	187	171	123	120	92	188	151	98	88	80	100	183	0	0	0	0	0	1	1	128	W
288	228	189	177	183	212	284	228	199	240	282	288	278	270	200	0	1	0	1	0	1	4	808	N
400	480	828	821	818	824	488	833	480	488	482	474	441	403	439	912	911	912	911	912	910	908	277	F

21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'	
13	10	17	14	18	20	32	39	82	88	89	89	81	2	3	1	0	0	0	0	1	1	5
148	79	110	97	88	100	98	100	87	73	47	89	128	1	1	1	0	1	0	0	0	128	W
284	230	182	188	181	188	138	180	170	177	200	189	189	0	1	0	0	0	1	0	333	N	
278	379	408	422	480	423	433	379	389	382	382	381	330	888	883	888	888	887	887	887	238	F	

24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'
2	8	10	24	13	14	28	28	39	27	18	31	48	87	88	48	0	0	0	0	0	1	2	2	5
38	88	17	100	72	88	13	19	100	70	89	13	10	18	80	70	0	0	0	0	0	0	0	7	W
143	84	89	48	88	89	71	79	83	87	128	98	188	170	183	87	1	0	0	0	1	1	0	288	N
224	241	292	208	288	237	289	282	178	284	188	289	184	188	139	232	407	408	408	408	407	408	408	141	F

20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'			
3	1	1	2	8	10	8	4	7	18	18	7	0	0	0	0	0	0	0	0	0	0	0	0
28	10	28	28	1	1	3	18	19	4	4	32	0	0	0	0	0	0	0	0	0	0	7	0
33	24	17	28	28	38	28	17	23	42	44	20	0	0	0	0	0	0	0	0	0	0	78	0
47	71	83	88	78	82	72	72	80	47	48	80	108	108	108	108	108	108	108	108	108	108	27	0

25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'
0	0	0	0	0	0	0	4	3	8	7	8	8	3	4	2	8	0	0	0	0	0	0	0	0	5
7	4	12	8	12	10	2	8	0	2	8	4	13	8	4	0	11	0	0	0	0	0	0	0	4	W
18	14	4	8	4	7	11	7	10	13	10	1	7	9	10	14	8	0	0	0	0	0	0	0	22	N
14	21	23	28	23	22	28	20	28	19	18	29	14	22	21	23	14	38	39	39	39	39	39	13	0	F

18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'
1	0	0	0	0	3	2	2	2	1	0	0	0	0	0	0	0	0	5
8	4	1	4	0	3	0	0	0	8	8	0	0	0	0	0	0	1	W
3	3	8	3	8	3	4	3	8	1	0	0	0	0	0	0	0	10	N
9	12	13	12	11	11	13	14	8	10	19	19	19	19	19	19	19	8	F

19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'
0	4	3	8	7	8	8	3	4	2	8	0	0	0	0	0	0	0	5	
1	2	3	2	0	0	4	2	2	0	8	0	0	0	0	0	0	0	W	
4	3	8	1	4	2	4	4	2	7	3	0	0	0	0	0	0	10	N	
11	11	7	12	10	13	8	8	9	7	8	18	18	18	18	18	18	2	F	

We initially built 1st order Markov models with n states dealing with different lengths of miRNA from 17 nt to 25 nt as shown in Figure 3(a). Each model deals with a pattern of sequences n characters in length, the same length as the corresponding miRNA. The training data was grouped by length and used to train the models separately. The number of states in the Markov chains was trained separately as well. Each state has four emissions representing the

binding status of corresponding nucleotides between the miRNA and its target. The binding status is the pattern of sequences converted from the Watson-Crick base pairs.

TABLE 3: EMISSION PROBABILITIES FOR THE 22 STATE 1ST ORDER MARKOV MODEL

	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	miRNA 5'
	46	46	46	60	81	125	140	132	122	140	144	210	260	275	3	0	0	0	0	1	2		4 B
	472	482	303	374	392	261	229	307	361	388	315	268	220	305	4	0	19	2	5	0	3		338 W
	779	473	407	384	506	546	512	458	505	467	482	516	635	586	0	9	4	3	0	4	2		1056 N
	835	1131	1376	1314	1153	1200	1251	1205	1144	1137	1191	1138	1017	966	2125	2123	2109	2127	2127	2125			734 P
	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1		
BB	0.55	0.71	0.67	0.57	0.46	0.50	0.50	0.43	0.54	0.50	0.63	0.60	0.39	0.25	0.22	0.11	0.17	0.25	0.25	0.25	0.40		
BW	0.10	0.07	0.13	0.29	0.08	0.14	0.07	0.07	0.08	0.08	0.06	0.07	0.06	0.06	0.11	0.33	0.17	0.25	0.25	0.25	0.20		
BN	0.10	0.07	0.07	0.07	0.08	0.21	0.07	0.07	0.08	0.08	0.06	0.07	0.11	0.06	0.11	0.11	0.17	0.25	0.25	0.25	0.20		
BP	0.25	0.14	0.13	0.07	0.39	0.14	0.36	0.43	0.31	0.33	0.25	0.27	0.44	0.63	0.56	0.44	0.50	0.25	0.25	0.25	0.20		
WB	0.04	0.05	0.09	0.14	0.08	0.25	0.14	0.15	0.07	0.09	0.07	0.13	0.10	0.15	0.08	0.17	0.08	0.20	0.25	0.20	0.11		
WW	0.28	0.26	0.13	0.07	0.08	0.13	0.14	0.39	0.27	0.18	0.27	0.13	0.10	0.08	0.25	0.17	0.08	0.20	0.25	0.20	0.11		
WN	0.12	0.26	0.09	0.36	0.17	0.25	0.29	0.15	0.07	0.09	0.13	0.13	0.20	0.15	0.17	0.50	0.08	0.20	0.25	0.20	0.44		
WP	0.56	0.42	0.70	0.43	0.67	0.38	0.43	0.31	0.60	0.64	0.53	0.63	0.60	0.62	0.50	0.17	0.75	0.40	0.25	0.40	0.33		
NB	0.05	0.06	0.06	0.05	0.11	0.07	0.05	0.07	0.12	0.10	0.05	0.11	0.07	0.11	0.29	0.17	0.10	0.20	0.20	0.14	0.14		
NW	0.18	0.25	0.12	0.37	0.07	0.10	0.09	0.15	0.08	0.14	0.18	0.16	0.29	0.11	0.14	0.17	0.10	0.20	0.20	0.29	0.43		
NN	0.36	0.31	0.41	0.26	0.46	0.39	0.41	0.33	0.39	0.29	0.50	0.21	0.36	0.11	0.14	0.17	0.20	0.20	0.40	0.43	0.14		
NP	0.41	0.38	0.41	0.32	0.36	0.45	0.46	0.44	0.42	0.48	0.27	0.53	0.29	0.68	0.43	0.50	0.60	0.40	0.20	0.14	0.29		
PB	0.03	0.06	0.02	0.04	0.09	0.07	0.08	0.07	0.02	0.13	0.07	0.11	0.13	0.02	0.06	0.04	0.01	0.01	0.01	0.03	0.16		
PW	0.20	0.27	0.17	0.24	0.09	0.34	0.18	0.12	0.09	0.17	0.16	0.09	0.13	0.16	0.01	0.09	0.03	0.01	0.02	0.06	0.13		
PN	0.10	0.13	0.21	0.34	0.41	0.11	0.28	0.33	0.21	0.26	0.11	0.15	0.18	0.04	0.03	0.07	0.01	0.02	0.04	0.03	0.25		
PP	0.67	0.54	0.60	0.38	0.41	0.48	0.48	0.49	0.67	0.43	0.66	0.66	0.56	0.78	0.90	0.80	0.94	0.95	0.93	0.89	0.46		

A difficulty of this topology is we need to switch between models according to the length of any given binding pattern sequence. Since most miRNAs are 22 nucleotides long, we have too few training data sequences shorter or longer than 22. Inadequate data causes accuracy problems for all probabilistic models, including Markov Models.

a.

b.

c.

Figure 3: Topologies of (a) the 1st order Markov Model and (b, c) hidden Markov Models

Therefore, we enhanced the simple 1st order Markov chain to a profile hidden Markov Model as shown in Figure 3(b). This design has 18 match states and 1 insert state. The training data covers the complete range from 18 – 27 characters. We broke each sequence into two parts: the first part contains the first 9 characters from 3' to 5' and the last part has the rest of the sequence. The first part aligns to the left and the last part which contains the significant seed region aligns to the right. We took the first 9 and last 9 columns as the match-state and the entire

center part as one insertion-state as shown in Table 4. In this case, we solved the problem of having too few sequences that could be used as training data for those miRNAs shorter or longer than the rest of the sequences. The disadvantage of this approach is it requires more computation time compared to the previous design.

Considering that the columns in the insertion-states have weak patterns, we tried the third design shown in Figure 3(c). All match-states are the same as the second model except it has no insertion-states. We trained this model using the same aligned multiple sequences by omitting the center columns in the insertion-state. This model is simpler than the first model and as effective as the second one.

TABLE 4: MULTIPLE BINDING SEQUENCES ALIGNMENT

miRNA 3'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
	N	P	P	P	P	N	N	N	N	P	N	P	P	P	B	P	N	P	P	P	P	P	P	P	P	P	N
	P	N	N	P	P	P	P	P	N	-	-	-	P	B	P	N	B	P	P	P	P	P	P	P	P	P	N
	N	W	P	P	W	N	N	P	W	-	-	-	-	W	P	P	B	B	P	P	P	P	P	P	P	P	P
	N	P	P	P	P	N	W	P	N	-	-	-	-	N	P	P	B	P	P	P	P	P	P	P	P	P	P
	P	W	N	P	P	P	P	P	P	-	-	-	-	-	-	-	-	-	W	P	P	P	P	P	P	P	P
	■1	■2	■3	■4	■5	■6	■7	■8	■9					I1					■10	■11	■12	■13	■14	■15	■16	■17	■18

2.2 Data

The results from this project depend significantly on the quality of the training sets, the rules defining complementarity, and the statistical approach. Therefore, the very first step is to correctly retrieve the miRNA sequence, the target sequence and the alignment sequence from databases. In this project, we used two miRNA gene and target databases called tarBase [9] and miRNAmap [10]. PACdb [11], RefSeqGene [12] and nt [13] were used as the search space.

Tarbase, from Diana Lab, has a manually curated collection of experimentally supported miRNA:mRNA interaction information for several species such as human, mouse, drosophila and plant [9]. We downloaded 128 human records from *TarBase_V4* in a flat file format which provides both miRNA:mRNA sequences and alignments without miRNA names or IDs.

miRNAMap is an integrated database storing both known and putative miRNA genes and their targets. The known miRNAs for mammalian genomes of human, mouse, rat and dog are obtained from miRBase[3] while the experimentally validated miRNA target sites are from a literature survey[10]. For putative miRNA target sites, three prediction tools were applied: miRanda, RNAhybrid and TargetScan. The three criteria for filtering out false positive predicted targets were: 1) predicted by at least two tools listed above; 2) target genes contain multiple binding sites; and 3) the target sites are accessible.

miRNAMap has a total of 2,330,545 human records. We analyzed these records and found 445 different miRNA genes. The advantages of this database are: it provides not only the sequences of miRNA, target and alignment but also miRNA gene names and the start/stop position of target sites. It also contains more sequences than tarbase. On the down-side, it contains a huge number of “bad” records such as miRNA or target sites with missing or duplicated sequences. We retrieved a total of 4367 records out of the entire database with no duplication and full information on miRNA:mRNA sequences.

2.3 Rank the computational predicted miRNA targets

This Markov model, trained with known miRNA target sites from tarbase, is capable of evaluating the diverse results predicted by different algorithms. Firstly, we align the predicted

target site on mRNA with the corresponding miRNA and convert it into our binding pattern sequence. Then we run the converted binding pattern sequence with the trained 1st order Markov Model. The latter gives a probability value that indicates the similarity of the input sequence to the training data family.

III. Human miRNA targets PREDICTION

Since there is no perfect computational miRNA target site prediction tool for animals, it is valuable to try and discover a new solution that enhances the accuracy of the contemporary prediction results. Here we present a new approach to predict target sites according to the given miRNA gene sequence. There are two main parts: part one produces roughly selected candidate sites, and part two filters out those candidates assigned low probabilities of being functional miRNA targets.

The work flow is shown in Figure 4. We had two schemes to search for potential target sites of a given miRNA sequence. The first scheme used BLAST tools from NCBI and *Mathematica* [45] from Wolfram research. It took the complement miRNA sequence 3' to 5' as a query sequence and blasted it against a selectable database. The settings of the stand-alone BLAST program were customized to fit our needs. Then we wrote a *Mathematica* notebook to analyze the output file from BLAST. The mRNA with the most hits was more likely to be a real target gene. We retrieved the potential target sites according to the *Mathematica* analysis and then converted the binding pattern sequences for use in the pHMMs. The advantages of this approach are speed and scalability. Since BLAST is an approximate algorithm which is often used to solve computationally intensive problems, it is much faster than the Needleman-Wunsch algorithm.

The second scheme was built to examine the first one in a relatively small database. It searched the 7-mer complement seed region of miRNA 3' to 5' and retrieved the 52 nucleotide long sequence that contained the match site at the 3' end. Each sequence produced 35 potential

target sites from 18 to 52 nucleotides in length. This was a faster algorithm avoiding the dynamic programming implemented in the Needleman-Wunsch algorithm. It covered all possible target sites for each potential seed region. The disadvantage of this approach is it is slow in searching a large database such as *nt*[13].

Figure 4: Project Work Flow

3.1 Blast Scheme:

The overview of this scheme is shown in Figure 5. For the given miRNA sequence in the 3' to 5' direction, we first convert it to the complement sequence which has the default orientation 5' to 3' used in mRNA sequence databases. The seed region is then at the 3' end. Because the length of the query sequence is 9 nucleotides, blast can find a large number of hits all with perfect alignments. However, for predicting animal miRNA target sites, perfect alignments are not suitable, especially at the 5' end of the mRNA which always contains some bulges and mismatches. It is almost impossible for blast to find hits containing mismatches for such short query sequences.

Figure 5: Blast Scheme Work Flow

3.1.1 Set up local blast and databases

BLAST is a standalone application provided by NCBI (National Center for Biotechnology Information). The BLAST tool `ncbi-blast-2.2.24+-x64-linux` from `ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/` was downloaded to a Dell server running an Xubuntu version 9.10 operating system. The advantage of using the local blast instead of web BLAST is we were able to construct our own customized BLAST database from sequences in FASTA format. In this project, we used three different BLAST databases: PACdb, RefSeqGene and nt.

PACdb is a database introduced by NCBI that contains the PolyA Cleavage Site and 3'-UTR sequences from human, mouse, fruit fly and *C.elegans* [11]. It is a flat file in FASTA format using ENSEMBL gene ids. This project used `Human_build36_vh.fa` which has 6051 records and 7,660,065 letters in total. Since *.fa file can not be used as a BLAST database directly, we had to use the program called *formatdb* to format the sequences prior to using them.

ReSeqGene is a FASTA format flat file downloaded from NCBI that contains genomic gene-region sequence data which is the foundation for medical, functional and diversity studies [12]. NCBI claims it is a reliable resource for gene recognition and classification. It has 3,435 records and a total of 285,825,825 letters.

nt is a ready-to-use pre-formatted database for BLAST. It holds nucleotide sequences from every one of the traditional divisions of GenBank, EMBL and DDBJ [13]. Therefore, nt is not a non-redundant data base that divided into 5 parts. Each part is approximately 2 GB. It follows, NCBI accession numbers to label each record.

3.1.2 Generate the BLAST query sequences from given miRNA

For an input query sequence as short as ~22 bases long, BLAST outputs only the perfect alignments since for shorter query sequences many exact matches exist in the database. We compared the results from the complement sequence of a single miRNA and its target site sequence. They were blasted against both PACdb and RefSeqGene and we took the top 500 hits each. There were no hits from the complement sequence result, because all of the hits in both files are almost perfect alignments or part of consecutive perfect alignments, whereas the majority of animal miRNA target sites contain 2-6 mismatches at the 5' end.

In order to conquer the perfect alignment problem, we tried different settings of the penalty and award scores for gap, mismatch and match but the problem persists because BLAST has restrictions on the penalty and award score setting. We changed the score through the command line but it did not help.

On the other hand, we tried to add single nucleotide mutations into the query sequence before submitting the query to BLAST. The first 9 nucleotides from the 5' end, the comparatively insignificant non seed region, often contain 2-6 mismatches and one bulge. If the sequence has n mutations, there are C_n^9 different permutations of mutation positions. For each permutation, there are 3^n possible mutated sequences. For each sequence, there are 9 possible single deletions. The number of possible mutated sequences containing n mutations, M_n for the first 9 nucleotides is:

$$M_n = C_n^9 * 3^n * (C_1^9 + 1) \quad \text{(Equation 1)}$$

$$\text{Two mismatch + One bulge: } C_2^9 * 3^2 * (C_1^9 + 1) = 3240$$

$$\text{Three mismatch + One bulge: } C_3^9 * 3^3 * (C_1^9 + 1) = 22,680$$

$$\text{Four mismatch + One bulge: } C_4^9 * 3^4 * (C_1^9 + 1) = 102,060$$

$$\text{Five mismatch + One bulge: } C_5^9 * 3^5 * (C_1^9 + 1) = 306,180$$

$$\text{Six mismatch + One bulge: } C_6^9 * 3^6 * (C_1^9 + 1) = 612,360$$

Therefore, the total number of all possible mutated sequences for the 9-base sequence at the 5' end with 2-6 mismatches and 1 deletion is the sum of M_2 to M_6 (1,046,520). The other 9-base sequence at the 3' end which contains the seed region only needs 1-2 mismatches at the 5' terminal and 1 mismatch at the 3' terminal. The query file that we submitted to blast contained all the mutated sequences generated from the given miRNA.

2.1.3 Run Blast

We used the blastn program to search nucleotide databases using nucleotide sequence queries. It was slower than megablast, but more sensitive with shorter default word size. We used the default word size of 11, along with the blastn-short option which is optimized for primer or short nucleotide motif (< 20 bases) searches[47]. We got consecutive perfect matches with the default match/mismatch score of +1/-3. The other parameter sets like +1/-1 and +3/-3 did not help with mismatches and gaps in the alignment. So we created a mutated sequence family from the complement sequence of given miRNA as the blast query file. Here is a typical command string to run the local BLAST program:

```
#./blastn -task blastn-short
        -query cputest.fna -db nt
        -out t.out
        -num_descriptions 0 -num_alignments 5
```

We changed the number of one-line-descriptions to 0 because alignments already included all information in one-line-descriptions, and smaller files are easier to parse. We experimented with

different values for the output sequence alignment, which defaults to 250, to satisfy different needs.

Although having all possible mutations in the query file may increase the chance for getting more hits on the target gene, the huge number of query sequences we generated heavily slowed down processing speed, especially when blasted against a large database such as *nt*[13]. The tradeoff was between accuracy and speed. One solution to this problem is parallel computing. Despite the fact that the *ncbi-blast-2.2.24+-x64-linux* we used does provide a parameter for the user to input the number of CPUs to use, it just switches between CPUs and never does real parallel processing. Fortunately, Perl is able to fork multiple processes and implement BLAST using parallel processors.. Here is a Perl example of forking using multiple processors:

```
for ($i=0;$i< NUM_OF_CPU;$i++) {
    $pid = fork;
    if ($pid == 0) {last;}
}

if ($pid == 0) {
    $j = $i + 1;
    print ("child $PID i = $i\n");
    exec ("time", "./blastn", "-task", "blastn-short", "-query",
        "cputest$j.fna", "-db", "nt", "-out", "t.out", "-num_descriptions", "5");}
```

We performed the parallel computing experiment on a Dell workstation that has 8 CPUs and 32 GB ram running Xubuntu 64-bit operating system against different sizes of blast databases. As shown in Figure 6, running a single BLAST process gets almost the full usage from one CPU. Figure 7 shows the CPU performance after splitting the input query file into 5 parts with an equal number of query sequences and running 5 BLAST processes in parallel, the 5 CPUs are nearly 100% busy. For those BLAST databases less than two gigabytes in size, the running time was almost 5 times faster.

Figure 6: CPU Usage Before Parallel Process

We tried making copies of the BLAST database for each process but it did not speed up the run time. This result indicates that each process may read the entire database into ram. As long as the ram space is adequate, every process runs independently.

Figure 7: CPU Usage During Parallel Processing

2.1.4 Analyze BLAST output

The BLAST query file that we construct has multiple sequences with annotations in FASTA format. Figure 7 shows the format of BLAST *.out file. The BLAST result for each query begins with “Query =”, followed by the query sequence annotation in the FASTA line in the query file. As shown in Figure 7, the annotation for the query sequence is “query 1”. Every hit for this query sequence starts with the FASTA annotation line of the subject sequence, which has “>” in front. This is where we retrieve the Sequence ID. The details of the BLAST hit follow, and the alignment is the last part of each hit.

```
Query= query 1
Length=18

> X:53423749:53419216:53419216:6:ENSG000000072501:53466274:53423749
Length=4533

Score = 36.2 bits (18), Expect = 5e-04
Identities = 18/18 (100%), Gaps = 0/18 (0%)
Strand=Plus/Plus

Query 1 CAACCTTCCCCCTACCTC 18
        |||||
Sbjct 74 CAACCTTCCCCCTACCTC 91

> 8:57085224:57089282:57089281:2:ENSG00000147507:56954949:57085223
Length=4058

Score = 26.3 bits (13), Expect = 0.47
Identities = 13/13 (100%), Gaps = 0/13 (0%)
Strand=Plus/Plus

Query 6 TTCCCCCTACCTC 18
        |||||
Sbjct 551 TTCCCCCTACCTC 563
```

Figure 8: Sample Blastn Output File from PACdb

According to the rule that each miRNA has multiple target sites on a target mRNA, we parse the BLAST output file to count the frequencies for each hit's sequence ID. Sequence data provided by different organizations such as GenBank, EMBL, DDBJ and PDB have unique styles of FASTA lines. Furthermore, they label genes in different ways.

TABLE 5: FASTA STYLE OF VARIOUS DATABASES

DB Name	FASTA LINE
refseqgene	> gi 225637498 ref NG_011690.1 Homo sapiens Alstrom
PACdb	> 22:39092473:39095316:39095315:2:ENSG00000100357
EMBL	>emb FN600646.1
GenBank	>gb GQ850526.1
DDBJ	>ref NG_005838.1
igSeqNt	> gi 2072142 gb U56414.1

Sequences that have multiple hits are likely to be potential miRNA target genes. We grouped the hits by sequence ID and calculated the distance between two hits inside each group. If a pair of hits satisfied the following three criteria:

- i) One hit from the first-part of the query sequence and the other from the last-part;
- ii) The hit's first-part position is prior the hit's last-part position; and
- iii) The distance between two hits is smaller than 26, since the longest human target site obtained from tarbase has 44 bases. ($44 - 9 - 9 = 26$).

Then we consider this pair of hits is coming from potential target sites. As introduced previously, the query sequences are mutated from the complement miRNA, therefore we need to convert them into the 18-base binding pattern sequence to be examined by the pHMM.

2.1.5 Feasibility Test

Figure 9: Feasibility Test Diagram

In order to test the presented method, we created some test cases from Tarbase_V4 with PACdb, RefSeq. Tarbase_V4 has 128 experimentally supported human miRNA:mRNA interaction records. We first retrieved the target site sequences from each record and put them into one single query file in FASTA format. We then blasted this file against PACdb and RefSeq to see if these target sites were able to map back into the mRNA sequences in those databases. Unfortunately, only ~10 out of 128 target sites have a single hit of 100% perfect alignment in both of the two databases. We imply from this result that TarBase_V4 has a completely different

germ line from NCBI and EMBL. Because of this germ line mismatch, we had no useful data to continue the test.

3.2 Seed Matching Scheme

This approach is suitable for relatively small databases such as PACdb and RefSeqGene other than the large ones like nt. First, it takes the 2nd to 8th bases from the given miRNA gene 5' to 3' as the 7-mer seed. Then it uses the reverse-complement sequence of the seed as one word search for exact matches through a specific database. This database should be a flat file in FASTA format. We analyzed both TarBase_V4 and miRNAmap, and the result shows human miRNA target site lengths range from 17 nt to 52 nt. A large number of miRNA:mRNA bindings have various sizes of loops in the centre. So we simply took all 35 possible targets sites for each seed match.

Once a match was found in an mRNA sequence, we took 44 bases in front and 1 base after the seed match site. This gave us a 52-base long sequence which may contain the potential target site from 17 to 52 nucleotides long. If this sequence is in the 5' to 3' orientation, we can print the sequence in FASTA format and cut the first base 34 times to obtain all 35 possible candidates. Every sequence only aligns the first and last 9 bases with the given miRNA as shown in Figure 10. These two parts are converted to the 18 bases binding pattern sequence and processed through the profile Hidden Markov Model. Unlike existing prediction algorithms, we didn't use RNA folding algorithms or calculate the free energy to identify the target sites. Instead, we let our profile hidden Markov Model process only the binding pattern sequence.

Figure 10: Single miRNA Seed Site has 35 Possible Target Sites

TABLE 6: CANDIDATE TARGETS SITES FOR A SINGLE 7-MER SEED MATCHING

>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 17 end 964 ATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 18 end 964 CATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 19 end 964 TCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 20 end 964 TTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 21 end 964 ATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 22 end 964 CATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 23 end 964 GCATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 24 end 964 TGCATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 25 end 964 TTGCATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 26 end 964 TTTGCATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 27 end 964 TTTTGCATTCATTTAATTTTAGGTCAA
>2:144863488:144858535:144858535:12:ENSG00000169554:144991387:144863488start 28 end 964 TTTTTGCATTCATTTAATTTTAGGTCAA

3.3 Hidden Markov Model

This Markov Model contains two identical Markov Chains. Each Markov Chain has 18 match states and each state has 5 emissions with no insertion or deletion state. One Markov Chain is trained with the real miRNA:mRNA interaction data and the other is trained with synthetic data to function as the negative model.

Figure 11: The pHMM Topology

3.3.1 Training the Model

We used four-fold cross validation to train the model. For the positive model, we converted 4000 refined miRNA:mRNA alignments from the miRNAMap into binding pattern sequences and divided them into 4 groups. Each group had 1000 sequences. Each time, we took 3 groups to train the model and tested it with the one group left out. We repeated this process 3 more times. The lengths of the binding pattern sequences in the training data varied from 18 to

27 which are the same as the lengths of miRNAs. We aligned the first 9 bases to the left and last 9 bases to the right, so that we only profiled the binding patterns at the two terminals because the centre areas vary a lot and have weak patterns.

The transition probability $a_{x_{i-1}x_i}$ for state S_i was calculated by Equation 2. $C_{x_{i-1}x_i}$ is the number of times letter x_i is followed by letter x_{i-1} in multiple sequence alignments. Due to inadequacy of training data, some values of $P_{(X|Y)}$ are zeros. Therefore, we added pseudo counts to our formula as shown in Equation 2. In order to avoid underflow problem, we used summation of log values instead of the product of probability. We used Equation 2 for the negative model.

$$a_{x_{i-1}x_i} = \frac{C_{x_{i-1}x_i} + 1}{\sum_{x_i'} C_{x_{i-1}x_i'} + 5} \quad (\text{Equation 2})$$

TABLE 7: ONE PROFILE CALCULATED FOR A SET OF POSITIVE TRAINING DATA

SS	414	148	459	345	395	325	223	190	533	332	109	197	135	85	349	443	923	
SB	1	178	1	1	214	1	1	142	43	143	17	20	81	10	4	1	1	
SW	242	1	1	236	1	1	274	1	123	82	58	99	76	22	50	28	4	
SN	1	1	187	1	1	197	1	1	66	49	113	5	58	112	33	13	2	
SP	402	484	156	421	229	275	553	231	271	526	493	273	540	361	252	531	118	
BS	1	1	221	1	1	259	1	1	78	66	123	68	22	50	42	5	1	
BB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
BW	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
BN	1	1	85	1	1	94	1	1	23	21	50	24	15	40	10	3	1	
BP	1	1	127	1	1	122	1	1	9	20	87	16	63	156	15	6	1	
WS	1	192	1	1	198	1	1	194	60	117	11	57	85	84	23	114	51	
WB	1	70	1	1	79	1	1	112	24	51	23	14	60	17	3	1	1	
WW	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
WN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
WP	1	173	1	1	178	1	1	187	26	85	95	17	81	21	72	126	4	
NS	178	1	1	190	1	1	136	1	125	13	29	88	15	32	112	37	12	
NB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
NW	84	1	1	90	1	1	89	1	52	35	17	55	25	24	50	9	3	
NN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
NP	180	1	1	163	1	1	224	1	49	54	58	117	76	49	74	9	2	
PS	218	462	322	303	204	466	204	404	336	262	322	480	333	437	490	449	847	
PB	1	185	1	1	182	1	1	232	40	66	68	66	105	40	7	1	1	
PW	109	1	1	129	1	1	130	1	78	12	13	72	21	52	141	19	10	
PN	1	1	171	1	1	158	1	1	13	34	97	87	32	84	12	1	1	
PP	182	117	281	132	330	113	176	317	70	51	236	264	196	343	280	223	36	
S	0.745	0.477	0.468	0.691	0.512	0.461	0.738	0.113	0.45	0.811	0.45	0.163	0.57	0.156	0.311	0.703	0.734	1.295
B	-6.22	-6.22	-0.15	-6.22	-6.22	-0.06	-6.22	-6.22	-0.04	-1.56	-0.66	-1.55	-1.63	-0.72	-2.04	-3.73	-6.22	-6.22
W	-6.22	-0.15	-6.22	-6.22	-0.1	-6.22	-6.22	-0.02	-6.22	-0.69	-1.37	-1.76	-0.81	-1.43	-1.65	-0.74	-2.23	-3.51
N	-0.13	-6.22	-6.22	-0.13	-6.22	-6.22	-0.12	-6.22	-6.22	-1.61	-1.59	-0.66	-1.48	-1.58	-0.76	-2.25	-3.51	-5.52
P	0.011	0.419	0.432	0.114	0.354	0.383	0.013	0.64	0.38	-0.17	0.379	0.655	0.309	0.641	0.614	0.318	0.575	-1.16

For the negative model, the same set of miRNAs used to train the positive model was aligned with random 3'UTR sequences and the random alignments were converted to binding pattern sequences. The results were full of Ns which represent non-pairs in every sequence. This

insured that any binding pattern sequence from a potential target site would be classified to the positive model, because all candidates have a consecutive seed region. So these random 3'UTR sequences were not suitable to train the negative model. Then we took the same miRNA sequences and converted them to binding sequences as if they all perfectly bind to the mRNA sites: A or U converts to P, which represents a Watson-Crick pair, and G or C converts to S, which is a strong binding. The problem was there was no mismatch, wobble or bulge in the sequence at all. So we randomly mutated n (where n is flexible) bases into N, W, B in the non-seed region.

TABLE 8: POSITIVE DATA V.S NEGATIVE DATA

ptrain1.txt		ntrain1.txt	
1	NWPSWNNPWWPWSBPSSPPSPN	1	SSPSSBNPSSPPSWPNPSPSPS
2	WWPSWNWPWWPPSNPSSPPSPS	2	NWPNWSSWSSPPSSPPSPSPS
3	NSPSSNWPNNPPBSPSSPPSPN	3	SWPSSBSPSNPPNWPSSPPSPS
5	SSPSNBBBWWPPSWPSSPPSPN	5	SSBSWSSPSSWPNNWPSSPPSPS
6	NWPSWNSPNSPPNNPSSPPSPW	6	NWPSSBSPBWPSSPPSPSPS
7	NWPSWNNPSSPPSWPSSPPSPN	7	SWPNSSSPSSPPNWPSSPPSPS
9	WWPSNSSWSWWPNWPSSPPSPN	9	SWPNSSSPSNPPNSPSSPPSPS
10	NNWSSSWPNPSPNPSSPPSPN	10	SSPSSBSPSNPPNSPSSBNSPS
11	WWPNNNSPNWPPSNPSSPPSPS	11	NSPSSBSWBSWPSSPPSPSPS
13	WWNSWNWPWWPPSWPSSPPSPN	13	SWPSWSSPBSPSSPNPSPSPS
14	NSPSNSBWWPSPNPSSPPSPN	14	NSPSSSSWSSWPSSPSSBPSPS
15	WSPSWNSNNWPSPWPSSPPSPN	15	SWPSSBSPSSWPNSPSSPPSPS
17	NNWSWSSPSWPPSWPSSPPSPN	17	SSPSSSNWSPNPSSPNPSPSPS
18	WWPSSSSPSWPPNSPSSPPSPN	18	SSBSWSSWSSPPSSPNPSPSPS
19	NNNNWSWPSNPPSWPSSPPSPN	19	SSPSSBSPSSWPNSPNSPPWPS
21	BWPSNNSPWPNNPSSPPSPN	21	SSPSSSNPSSPPSWPSSPNWPS
22	NNPSWNWPBPPSNPSSPPSPS	22	SSBSSSPSSPPSWBSBPSPS
23	SSWSSSWPNWPBBPSSPPSPS	23	SSBSSBSPSSPPSSBSSBPWPS
25	NNNNWNSWSSPPNNPSSPPSPN	25	SWPSSSSPSSWPNSBSSPNSPS

3.3.2 Running the Model

When both of the positive and negative models are trained, we tested with the group left out from both the positive and negative data. We processed one sequence a time: given a sequence, we computed the log-odds ratio shown in Equation 3 to discriminate between the two models [50]. $S(x) > 0$ means x is likely to belong to the positive model, which means the corresponding target site has high probabilities to be a real one.

$$S(x) = \log \frac{P(x | \text{the + model})}{P(x | \text{the - model})} = \sum_{i=1}^L \log \frac{a_{x_{i-1}x_i}^+}{a_{x_{i-1}x_i}^-} \quad (\text{Equation 3})$$

TABLE 9: FOUR-FOLD CROSSAPPENDICES

		Positive Data		Negative Data	
		True Positive	Ture Negative	False Positive	False Negative
PROFILE 1	p test 1	98.5%	1.5%	n test 1	0.2%
	ptest 2	98.7%	1.3%	ntest 2	0.2%
	ptest 3	98.5%	1.5%	ntest 3	0.2%
	ptest 4	99.0%	1.0%	ntest 4	0.2%
PROFILE 2	ptest 1	98.6%	1.4%	ntest 1	0.2%
	p test 2	98.9%	1.1%	n test 2	0.1%
	ptest 3	98.5%	1.5%	ntest 3	0.0%
	ptest 4	98.8%	1.2%	ntest 4	0.1%
PROFILE 3	ptest 1	98.6%	1.4%	ntest 1	0.2%
	ptest 2	99.0%	1.0%	ntest 2	0.1%
	p test 3	98.5%	1.5%	n test 3	0.0%
	ptest 4	98.8%	1.2%	ntest 4	0.1%
PROFILE 4	ptest 1	98.6%	1.4%	ntest 1	20.0%
	ptest 2	98.9%	1.1%	ntest 2	10.0%
	ptest 3	98.4%	1.6%	ntest 3	0.0%
	p test 4	98.8%	1.2%	n test 4	0.1%

III. CONCLUSION

Imply from the four-fold-cross validation result, we have average 98.7% true positive rate and 99.9% false negative rate for the 1000 sequences from the positive and the negative training data. This result proves that miRNA:mRNA interactions do have a binding pattern, and our 1st order profile hidden Markov Model is effective to recognize the pattern. It seems to be a promising way to examine RNA folding/binding without actually calculating free energy, because we differentiated G:C triple hydrogen bond from A:U double hydrogen bond, G:U wobble pair, non Watson-Crick pair and bulge in the binding pattern sequence.

Although the profile hidden Markov Model works fine with four-fold-cross validation, it passed all 35 candidate sequences generated by our Seed-Matching approach, which should only pass the one real sequence and fail the other 34 sequences. This problem was caused by the synthetic training data for the negative model. If more experimentally verified target sites were able to map back to the databases, we could train the negative model with the 34 wrong predicted target sites. Unfortunately, there are only <10 target sites from TarBase_V4 found in the database.

VI. FUTURE WORK

It is essential to build a database for the experimentally confirmed false predicted miRNA gene target sites. The negative training data is as important as the positive data.

It is unfortunate that this project did not result in a positive prediction of target sites for miRNA. However, it does show the flexibility of the pHMM architecture, and the promise that such architectures hold. It is sincerely hoped, that when enough data does become available, pHMMs as suggested herein, can produce useful predictions of miRNA target sites.

REFERENCES

- [1] Lai, E.C.(2003), microRNAs: Runtts of the genome assert themselves, *Curr. Biol.*, 13, R925-R936
- [2] Watson JD, Crick FH (1953), Molecular structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid, *Nature* 171 (4356): 737–738.
- [3] http://www.mirbase.org/cgi-bin/mirna_summary.pl?org=hsa
- [4] Sean R. Eddy, Profile hidden Markov models, *BIOINFORMATICS REVIEW*, Vol 14 no. 9 1998, 755-763
- [5] David P. Bartel, MicroRNAs: Genomics, Biogenesis, Mechanism, and Function, *Cell*, January 23, 2004 by Cell Press, Vol. 116, 281-297
- [6] Zvi Bentwich, Micro RNAs play a central role during cancer and viral infections and can serve as targets for therapy, *Retrovirology*(2006)
- [7] N. D. Mendes, A. T. Freitas and M.-F. Sagot, Current Tools for the identification of miRNA genes and their targets, *Nucleic Acids Research*, 2009
- [8] Wiemer, Erik A C C. "The role of microRNAs in cancer: No small matter." *European journal of cancer* 43.10 (2007): 1529-1544
- [9] Praveen Sethupathy, et al. TarBase: A comprehensive database of experimentally supported animal microRNA targets, *RNA* 2006. 12:192-197
- [10] Paul W.C. Hsu, et al. miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Research* 2005, vol 34, D135-D139
- [11] Brokman JM, et al. PACdb: PolyA Cleavage Site and 3'-UTR Database. *Bioinformatics*. 2005 Sep 15;21(18):3691-3
- [12] ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/
- [13] <ftp://ftp.ncbi.nih.gov/blast/db/README>
- [14] Grad Y et al, "Identification of many microRNAs that copurify with polyribosomes in mammalian neurons" *Mol. Cell.* (2003);11:1253–1263.
- [15] Lim L.P. et al, "The microRNAs of *Caenorhabditis elegans*", *Genes & Dev.*(2003) 17: 991-1008

- [16] Pang K.C. et al. "RNAdb 2.0—an expanded database of mammalian non-coding RNAs", *Trends Genet.* (2006);22:1–5.
- [17] Pfeffer S., et al(2005), Identification of microRNAs of the herpesvirus family, *Nat. Methods.*;2:269–276.
- [18] Bentwich I. et al(2005), Identification of hundreds of conserved and nonconserved human microRNAs, *Nat. Genet.* 2005;37:766–770.
- [19] Vasudevan, S. et al, AU rich element mediated upregulation of translation by FXR1 and Argonaute, *Science* 2007;318:1931-1934
- [20] Ambros, V. et al, A uniform system for microRNA annotation, *RNA* 2003;9:277-279
- [21] Mendes, N. et al., Current tools for the identification of miRNA genes and their Targets, *NAR* 2009:1-15
- [22] Brennecke, J. et al., Towards a complete description of the microRNA complement of animal genomes, *Genome Biol* 2003;4:228
- [23] Sheng, Y. et al. Lenhard B. Mammalian microRNA prediction through a support vector machine model of sequence and structure, *PLoS ONE* 2007;2:e946
- [24] Helvik, S.A. et al. Reliable prediction of Drosha processing sites improves microRNA gene prediction, *Bioinformatics* 2007;23:142-149
- [25] Lin, L.P. et al., The microRNAs of *Caenorhabditis elegans*, *Genes Dev.* 2003;17:991-1008
- [26] Bentwich, I. et al, Identification of hundreds of conserved and nonconserved human microRNAs, *Nat. Genet.* 2005;37:766-770
- [27] Starck, A. et al., (2003) Identification of *Drosophila* MicroRNA targets, *PLoS Biol.*, 1, E60
- [28] Enright, A.J. et al., (2003) MicroRNA targets in *Drosophila*, *Genome Biol.*, 5, R1
- [29] John, B. (2004) Human MicroRNA Targets, *PLoS Biol.*, 2, e363.
- [30] Lewis B P, Shih I H, Jones-Rhoades M W, et al. Prediction of mammalian microRNA targets. *Cell*, 2003, 115: 787-798.
- [31] Farh KK, 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.
- [32] Rajewsky, N. and Socci, N.D. 2004. Computational identification of microRNA targets. *Dev. Biol.* 267: 529–535.

- [33] Rehmsmeier, M., Steffen, P., Hochsmann, M., Giegerich, R. 2004 Fast and effective prediction of microRNA/target duplexes RNA 101507–1517
- [34] Krek A, Grun D, Poy MN, et al. Combinatorial microRNA target predictions. Nat Genet 2005 Krek, A. et al. (2005) Nat. Genet. 37, 495–500.
- [35] Burgler, C. et al. (2005) PM 2005. Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method, BMC Genomics, 6, 88.
- [36] Gaidatzis, D. et al. (2007) Predicting miRNA profiles from AGO-immunoprecipitated transcripts, BMC Bioinformatics, 8, 69.
- [37] Saetrom, O. et al. (2005) Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA, RNA, 11, 995–1003.
- [38] Lim, L.P. et al. (2005) Microarray analysis shows that some microRNAs down regulate large numbers of target mRNAs, Nature, 433, 769–773.
- [39] Kertesz, M. et al. (2007) The role of site accessibility in microRNA target Recognition, Nat. Genet., 39, 1278–1284.
- [40] Luciano, D.J. et al (2004) RNA editing of a miRNA precursor, RNA, 10, 1174–1177.
- [41] Robins, H. et al. (2005) Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs, Proc. Natl Acad. Sci. USA, 102, 15557–15562.
- [42] Cai, X. et al. (2004), Human microRNAs are processed from capped polyadenylated transcripts that can also function as mRNAs, RNA, 10, 1957–1966.
- [43] Erno Wienholds, et al. (2005), MicroRNA function in animal development, FEBS 29869
- [44] Julius Brennecke, et al. (2005), Principles of MicroRNA-Target Recognition, PLoS Biol 3(3): e85.
- [45] Hao-Yue Fu, et al. (2009), Assessing potential miRNA targets based on Markov model. Genetics and Molecular Research 8 (3): 848-860
- [46] <http://www.wolfram.com/mathematica/>
- [47] http://www.incogen.com/public_documents/vibe/details/NcbiBlastn.html
- [48] Ian Korf, et al (2003), BLAST, O'REILLY, ISBN: 0-596-00299-8
- [49] Nikolas Rajewsky (2006), microRNA target predictions in animals, Nature Genetics Supplement, Vol 38
- [50] Sami Khuri, Biology_CS_123B_FOUR_Markov_2perpage.pdf 2002

APPENDICES

Script Name	i/p	o/p	DataBase
bindpatnConv.pl	ndtar3c.txt	Bindpattern.txt	miRNAMap
	Converts the miRNA:mRNA into binding pattern sequence		
bindpatnConvTarbs.pl	Pic_Human_Table.txt	tbsePattern.txt	tarbs_v4
	Converts the miRNA:mRNA into binding pattern sequence		
datafiletest.pl	miRNA_targets_hsa.txt	Ndtar3c.txt	miRNAMap
	Eliminate duplicated records		
migene.pl	Data.txt	mirnageneWTa.txt	miRNAMap
MMcutoff.pl	MMprofile.txt	Values for each i/p seq	
	tbsePattern.txt		
MMProfile.pl	tbsePattern.txt	MMprofile.txt	
	Calculate transition and emission probabilities in log value with pseudo counts		
MMrank.pl	MMprofile.txt	#seq, #passed, pass rate	
	Bindpattern.txt		
patnAnaly22only.pl	Rank input Seqs with pre-determined cut off value		
	bindpattern.txt	Pwmcaltxt(?)	miRNAMap
patnAnaly.pl	Produce PWM		
	tbsePattern		Tarbase_v4
query_gen.pl	Produce PWM		
	Mirnagen.txt	Has_mirna.fna	
Query_gen01.pl			
	mirnagene.txt	Has_mirna_comp.fna	
Query_gen02.pl	Yields query file in FASTA format for blast		
	Mirnagen.txt	Has_mirna_comp_seedmask1.fna	
Query_gen.pl	Masked non-seed region in lower case		
Tarbs_tar_only.pl	Pic_Human_Table.txt	Tarbs_tar.fna	Tarbs_v4
		Tarbs_tar_rev.fna	
RandBindpathConv.pl	Generate the query file forward and reverse		
	ndtar3c.txt	RandBP.txt	miRNAMap
Preselect.pl	rand3UTRsite.txt		
	Generate the negative training data from random 3UTR region		
Neg_seq.pl	miRNA	Tar3utr.txt	
	testdb.txt		
rand3UTR.pl	Scan database with x-mer seed to preselect candidate sites		
	newBindpattern.txt	ptrain1.txt	miRNAMap
		pptest1.txt	
	Create 4-fold cross validation data		
	has_3utr.fna	rand3UTRsite.txt	PACdb
	Out put random cut 3'utr sequences		