FEATURE EVALUATION OF THE SUPPORT VECTOR MACHINE FOR MICRO-RNA TARGET PREDICTION IN ARABIDOPSIS THALIANA BASED ON ANTISENSE TRANSCRIPTION AND SMALL RNA ABUNDANCE

by

VIKTORIA GONTCHAROVA

A MASTERS THESIS

IN

COMPUTER SCIENCE

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

MASTER OF SCIENCE

Dr. Eunseog Youn
Committee Chairman

Dr. Chris Rock

Dr. Richard Watson

Dr. Fred Hartmeister
Dean of Graduate School

December, 2007

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Micro RNAs (miRNAs) are small non coding RNA that contribute to post transcriptional regulation. They are 21-23 nucleotide long sequences that affect development by binding by Watson-Crick pairing to a target gene and antagonizing various pathways of expression. This thesis explores the miRNA binding within the *Arabidopsis thaliana* genome as it related to antisense transcription of target genes.

Presented is a prediction mechanism that is based on two related features, antisense transcription and small RNA abundance, hypothesized to be of the miRNA binding site in the target gene. A newly discovered phenomenon in antisense strand of the target genes was implemented as a novel feature for target gene prediction. This feature, along with small RNAs and a commonly used indicator of binding sites were used in a Support Vector Machine to build a prediction model.

The three features were incorporated and analyzed using the output of the Support Vector Machine. Comparison was made between predicted and validated classifications to evaluate the importance of the features. Based on the accuracy, specificity, sensitivity and precision of the SVM results, the newly discovered feature may be able to identify new miRNA target sites in *Arabidopsis* and other species with deep genomic resources.

CHAPTER 1

INTRODUCTION

There are few people in history who have had the opportunity to change the way future generations see the world. Although there is constant progress in the development of thought and knowledge in numerous disciplines of study, it is rare that something comes along that becomes the basis of all future study in the field.

One such idea was formed by Francis Crick, an English molecular biologist, physicist and neuro-scientist who contributed to research within the field of genetics. His collaboration with James D. Watson has led to the discovery of the structure of DNA. Their work with Maurice Wilkins also earned them a Nobel Prize for Physiology or Medicine “for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material.” If that wasn't enough, Crick also is responsible for naming a now well-known cellular process, the “central dogma” of molecular biology [7].

The Central Dogma of Molecular Biology explains the unidirectional synthesis of proteins from DNA. DNA is first transcribed into RNA and then translated into proteins. Although at the time the hypothesis was deemed a dogma, the name seemed appropriate. As Crick explained, he, “simply applied it to a grand hypothesis that, however plausible, had little direct experimental support” [8]. After decades of research within the field, the dogma still stands. The process of transcription and translation, as well as its subjects, DNA, RNA and proteins have been continually studied.

Great strides have been made in the field. Consider the fact that the entire genomes of species have been mapped, countless proteins have been discovered and their function understood. However, much more is left unknown. Within the genome of any species, a large percentage of the DNA is named non-coding DNA, although the name is a bit misleading. It is not necessarily true that the non-coding DNA doesn't have "instructions" for making proteins; it could be that the purpose of these parts of DNA has not been discovered yet. This idea also carries over to RNA. As mentioned before, DNA is transcribed into RNA, and the non-coding regions also exist. However, unlike with DNA, the connotation of "useless" information is not always as quickly placed.

Non-coding RNA genes are an umbrella under which numerous types of small RNAs, more particularly micro RNA (miRNA) genes lie. These structures have a hairpin-like shape and are constantly being discovered and confirmed. The miRBase, a database containing over 5000 miRNA gene loci continues to grow [14, 15].

MiRNAs are homologous to the reverse complement of portions of another gene's mRNA transcript. Furthermore, they are single-stranded molecules of about 21-23 nucleotides that contribute to post-transcriptional regulatory activity. The genes coding for miRNA are first transcribed as primary-miRNA, then processed to about 70-nucleotide stem-loop pre-miRNA in animals; in plants the pre-miRNAs are more heterogeneous, ranging from about 70 to 300 nucleotides. Mature miRNAs then develop in the cytoplasm in animals and in the nucleus in plants by interaction with the endonuclease Dicer. Subsequently the RNA-induced silencing complex (RISC) is programmed by incorporation of miRNAs, which provide specificity in post-transcriptional or translational repression of target mRNAs [2].

Although miRNAs are known to contribute to gene silencing within cells, it is the RISC complex that is actually responsible for RNA interference and gene silencing. RISC processes also extend to include DNA and chromatin modifications, although the mechanisms are poorly understood. After the Dicer cleaves the pre-miRNA hairpin, one part of the strand is integrated into the RISC complex, while the other strand is degraded as a RISC complex substrate [12]. After integration into the RISC, the miRNA forms Watson- Crick pairs with the complementary mRNA molecule and the specific effects of miRNA or RISC begin. Finding these target sites within mRNA is often the goal of bioinformatics researchers. Without the target site, the functionality of the miRNA cannot be determined; therefore the task of understanding what the miRNA does cannot be accomplished. A major obstacle to animal miRNA target prediction is the degeneracy of base pairing: as few as 7 out of 21 nucleotides are complementary to target mRNAs [16].

Considering the possible difficulties and obstacles that occur during the process of miRNA target site prediction, great strides have been made in the field. Luckily, a property in plants was discovered that allowed for hundreds of miRNA targets to be found. Unlike animals, plants exhibit an almost perfect complementarity between miRNAs and their targets. Although research within the miRNA field did not start with plants, the almost perfect complementarity spurred on the development. Initially about 150 miRNAs were discovered [30, 38]. However, after miRNA presence was confirmed within the plant kingdom [34], researchers were able to find scores of miRNAs and hundreds of target genes before going back to the animal kingdom. Now, years later, the research is continuing within both plants and animals.

As miRNA continues to prove its importance, evidenced by conservation across phyla, it is being more and more understood. The numbers within the miRNA data bank continue to grow despite the difficulties in identifying these small molecules. This growth is due in part interdisciplinary study. Although miRNA and their target sites exhibit certain characteristics that in theory could be seen through visual and biological analysis on a case-by-case basis, that process is slow and unreliable. It is difficult to analyze the huge arrays of data without some involvement of computers and other tools that not only automate the procedure, but also make it more accurate and amendable to statistical analysis.

Inter-field collaboration has eased the work of the scientists and produced results that otherwise could not have been attained. Through the years, algorithms have been developed that take advantage of the traits that miRNA and their target sites display and have produced candidate miRNAs that later have been confirmed. This thesis will explore such a task. In the past, complementarity of miRNA and the targets has been heavily exploited, drawing on free energy, or enthalpy, a useful tool for predicting the binding sites for miRNA. It has become a widely used characteristic within this sort of analysis. However, other predictors may exist. Dr. Chris Rock, at Texas Tech University, found another possible quality of miRNA and miRNA targets: their effects on antisense transcription of target mRNAs. Using this newly discovered idea and massively parallel signature sequencing [25] data of small RNA counts abundance the target site, along with the already established idea of free energy, a Support Vector Machine has been built that analyzes the importance of each of the characteristics, or features, while analyzing the ability to predict possible miRNA sites. With this knowledge, not only can a list of

possible target sites be generated in sequence genomes, but down the line, it may lead to discovery of new miRNAs and insight into novel functions or mechanisms. Also, although the analysis and study of the Support Vector Machine (SVM), is done on a plant, *Arabidopsis Thaliana*, the tool could possibly be used on other genomes including animals. If the newly discovered feature transitions between kingdoms, it would be a valuable asset for scientists in continuing to identify miRNAs, their targets, their mechanisms of action, and hopefully furthering our understanding of ourselves, and the world around us.

CHAPTER 2

RELATED WORK

This chapter explores the history of miRNA, target site discovery and the progress that has been made within the field over recent years. Beyond the topic of miRNA, this chapter will also explore some of the tools that are already available for miRNA target site prediction. Following these topics, the Support Vector Machine and feature evaluation will also be addressed.

2.1 Historical Contributions

Although genetic makeup of various genomes has been studied for years, it is relatively recently that miRNA has become a matter of interest. The first micro RNA was discovered genetically as a mutant gene in the nematode *Caenorhabditis elegans* [19] that influences the larva by affecting the timing of development. Lin-14 encodes is a protein that has been shown by analysis of mutations in both lin-4 and lin-14 to be negatively regulated by lin-4. After extensive and intensive lab work involving cloning and mapping of various lin-4 homologues from several species, it was concluded that the lin-4 genes all encode similar products. However, further research into the products of lin-4 resulted in observations that contradict the idea of lin-4 producing a protein, but a small non-coding RNA instead. This research, although revolutionary at the time, was somewhat under-appreciated. For quite some time, the non-coding activity witnessed in lin-4 was attributed to only be a phenomenon seen in worms.

It was the work of Reinhart *et al*. that convinced scientists that miRNA is not just a worm oddity. This group also worked on the C. *elegans* genome. However, the miRNA gene in question was let-7. After some testing, the researchers came to the understanding that let-7 maybe be responsible for the temporal switch between larval and adult fates [33]. Let-7 contributes to expression of LIN-29 in the L4 stage, which in the end is responsible for the fate of adult cells. It appears that the effects of let-7 are seen later than that of lin-14 so the pathways are active at different times. However, it is not only believed that lin-14 eventually affects the let-7 pathway, but let-7 can directly regulate lin-14 as well. Also, findings suggest that it is unlikely for let-7 to be translated, and it functions as an RNA molecule. There is also support that let-7 has complementary sites in lin-41 and when bound in the L3 and the adult stages, down regulation of lin-41 gene activity is seen. This interferes with the transition to the adult stage.

After this research, it was difficult to argue with the importance of miRNA, but further conclusions were made when it was found that Dicer-1 is an essential component for development as well. It was already established by several papers that Dicer1 enzyme is responsible for the production of miRNAs; however, Murchison *et al.* confirmed the importance of Dicer1 in vertebrate development [27]. Zebra fish with mutant phenotypes were studied against the fish with a normal phenotype and it was concluded that the pleiotropic mutant phenotype was caused by a Dicer-1 disruption. It was speculated that development through various stages was effected due to the failure to produce not only miRNA but also other small RNAs. Although his study had certain limitations, it was enough to demonstrate the need for more research and time to be allocated to this field of study.

2.2 micro RNA and micro RNA Target Prediction

Great strides in small RNA research have been made within the plant kingdom. For example, RNA interference was discovered first in plants and called post-transcriptional gene silencing, (PTGS) [12, 28, 40]. Further, the original mutants for micro RNA metabolism and action; DCLI and ARGONAUTE family members were discovered in the model plant Arabidopsis [9, 36, 4, 23]. Computational approaches were employed which not only made the work less tedious, but more efficient and accurate. One of the more influential papers on the topic was published by Rhoades *et al*. They contrasted miRNA characteristics between plants and animals and made some important conclusions that remain valid. The work began with finding mRNA complements with less than 4 mismatches to any 1 of the 16 new miRNA families that were found in *Arabidopsis* and rice. The control group consisted of permuted sequences of the same size and base composition as that of the miRNAs. When the constraint of having less than 2 mismatches was placed on the algorithm, results showed 30:2 hits to authentic miRNA [35]. On tests with less stringent criteria, the complementarity was still evident. The study was done on both the *Arabidopsis* and the *Oryza* genomes and it was found that the miRNAs are conserved between the two species. Many of the potential targets carry over from one genome to the other. Interestingly, like in worms, the predicted targets were predominantly transcription factors that are responsible for floral and meristem development. The processes involved in cell division and cell fate. However, one big difference between the two kingdoms is that quality of near perfect complementary in plants between the miRNA and the targets. Also, the miRNA complementary sites in plants are in the ORF region of miRNA, while in animals it is usually in the 3' UTR

region. There are also typically multiple sites in animal mRNAs, while in plants there are only a couple of examples out of hundreds. All this lead the scientists to conclude that plant miRNA resembles siRNA target recognition. This was an important possibility because if there is a resemblance in recognition, perhaps it also exists in action. Plant miRNA, unlike animal, might be able to target any region of the mRNA.

Following the idea of near perfect complementarity between miRNA and the target sites, research continued. Wang *et al*. pursued the topic and questioned the extent of conservation between *Arabidopsis* and *Oryza sativa* miRNAs [41]. The research began with the task of defining sequence and structural properties of the *Arabidopsis* miRNA. The miRNA sequences were attained from random genomic data received from scans for candidate miRNAs in the intergenic region in the *Arabidopsis* genome. The properties to be analyzed were decided to be:

1. Presence of the hairpin
2. G + C content
3. Hairpin – loop length in the precursor RNA structure
4. Number and distance of mismatches in the stem
5. Phylogenetic conservation of mature miRNA in *O. sativa*
6. Complementarity

The automated process resulted in 95 candidate miRNA genes, of while 12 families were already known and published. Of the 83 newly discovered candidates, 12 additional families were validated by a Northern blot test and 8 new ones were added to

the ASRP database.

To find the miRNA targets, prediction was done by aligning the miRNA sequences with the targets using the Smith-Waterman nucleotide-alignment algorithm. Gaps were allowed, but mismatches were preferred, as demonstrated by a higher penalty for a gap. The algorithm resulted in an average of about 48 putative targets per miRNA.

An interesting speculation is that many of the miRNA/target pairs are responsible for transcription regulation. Although that doesn't seem to be anything new, one possibility is that the regulation is plant or tissue specific. After analysis of the expression patterns of potential targets, it was found that 12 of the 14 miRNAs confirmed by the northern blot showed evidence of contributing to flower-specific regulation. The expression was very different from that of the general expression pattern of the micro array data. Using this idea, some of the target genes were identified by nearly perfect complementarity. However, due to the fact that whole-genome tissue comparison data was limited in availability, the experiment also showed limited, yet promising results.

2.2.1 Tools for miRNA Target Prediction

It quickly became clear that finding miRNAs and their target sites is not an easy or fast task to accomplish. Due to this, various tools began popping up with different approaches of tackling the problem of enormous amount of data and even larger amounts of possible answers.

One of the results of this dilemma was a tool called miRanda [9]. It was developed at the Computational Biology Center of Memorial Sloan-Kettering Cancer

Center. It was written in C and is now available as an open-source method. Two files are the inputs where the sequences of file 1 are scanned against those in file 2 for potential target sites. Identification of the sites is done as a process. It begins with a dynamic programming local alignment procedure testing for sequence complementarity. Next, the alignments with the high scores are analyzed for thermodynamic stability. A potential hairpin is formed with the two sequences connected by a linker and a Z-score is produced evaluating it.

Another product of the need for efficient and accurate target site prediction tools was Target Scan. This tool predicts targets by searching for conserved 8mer and 7mer sites that match the "seed region" of each miRNA. The seed region is comprised of nucleotide positions 2-7 of a mature miRNA which are nearly always perfectly conserved among the members of the miRNA family. Lewis *et al*. [20] used the over-representation of conserved adenosines flanking the seed complementary sites in mRNAs as the criteria for base pairing to indicate target sites. Their efforts proved the presence of targeting in the 3' UTR regions along with the open reading frames. Evidence showed that over a third of human genes have conserved miRNA targets.

Another one of the many tools available is RNAhybrid. Developed in 2004, it has shown to be a very effective algorithm for miRNA target site prediction. The quality used here is the minimum free energy hybridization of two strands [21]. Multiple potential binding sites of miRNAs are found in large target mRNAs. Energy is used to evaluate the best hybridization of small RNA to large RNA with no intra molecular hybridization being allowed. The optimal, suboptimal and non-overlapping energy favoring hits are the result. The algorithm has linear time efficiency proving to be a good tool to use in a

situation when large sequences are being analyzed.

Another tool using the string complementarity to its full advantage is miRU [43]. This is a server-based mechanism where the user inputs the mRNA dataset for the organism. A Perl script is used to complete and exhaustive sequence similarity search that is similar to the BLAST algorithm. The mismatches defined by G: U wobble pairs, insertions/deletions, and all other non-canonical Watson-Crick pairings can be limited, and the alignment score is the resulting answer. To further the accuracy, another search can be done with homologous miRNAs against the mRNA dataset to extract information about conserved homologous miRNAs and targets in different species. The system would then compare the two sets of results to determine if the homologous genes are predicted in both genomes. The output consists of the summary of search input parameters, a list of predicted complementary targets displayed in order of mismatch scores, and the target gene sequences in FASTA format with the target site highlighted. The tool seems to be effective since it was able to predict not only those targets identified by two other, very reliable studies, but found potential sites that the studies missed.

One of the more recent contributions to the ever growing need to find miRNA targets is a tool called miTarget. This is a support vector machine classifier using three classes of features. The features fall into the structural, thermodynamic and position-based qualities of the sequences [18].Although the results were compelling of the testing of miTarget, the performance could be improved with more training examples. Support vector machine analysis for miRNA of animals contributed to exciting results with prediction of 3 human miR genes and demonstrated importance of certain positions in the miRNA. The work of Kim *et al*. could be viewed as a template for the goal of this thesis.

2.3 Support Vector Machine

Support Vector Machines (SVMs) have an array of complexity associated with them. Christopher Burges explored this topic in 1998 in great detail and extended proofs in his tutorial of the SVM [5]. The main idea of the Support Vector Machine is to classify information. An example Burges uses is that of a pixelated image. If an image composes a tree, it is given the classification of +1, if there is no tree it is in the other category, -1. The Support Vector Machine needs to analyze the pixels and be able to organize the information in such a manner that a barrier could be formed between the two categories. This is done by analyzing the characteristics (features) of a tree. There need to be a number of them, so the machine can learn how to predict if an input picture has a tree or not. It is important to have a balance of features though that accurately can teach the machine to predict. For example, having the tree being green as the only feature would not be a very good learning system. However, having 100 more features that include precise information like number of leaves wouldn't be good either, since not every tree has the same number of leaves, and some have none at all. The goal is to teach a balance to the machine so it can discern between a tree and a non-tree. The learning of the machine takes place when the features along with a classification are provided. The machine learns what kind of features contribute to a tree, along with what features are lacking that are responsible for the picture being a non-tree. After the learning has taken place, a model is built. Based on the model, when new input is added, without the classification, the machine should be able to discern, based on the features if it is a tree or not. Of course the idea can be applied to not only picture identification, but any task where characteristics can be converted to a numerical value. Within this thesis, gene

expression levels from micro-array, small RNA counts and free energy of the sequences will be the three features under analysis of the SVM.

2.4 Feature Selection and Evaluation

It is evident that feature selection is an important topic within the realm of target prediction and Support Vector Machines; after all, the analysis has to be made in respect to some characteristic, and it needs to be relevant to make a valid prediction.

The goal of feature selection is to increase efficiency and maintain accuracy of a mechanism. Because of this, it is necessary to not only select relevant features, but also eliminate irrelevant ones and combine those that are left in an effective manner. Blum and Langley explore this idea in their book on feature selection [3]. One of the earliest and most common feature selection mechanisms known is the nearest neighbor approach. Here, the test instances are classified by retrieving the nearest stored training example. Although it is an easy and logical approach, the irrelevant attributes have a great potential of slowing down the learning. Also, the number of training examples has potential to be huge, since they grow at an exponential rate with the number of irrelevant examples.

Due to this fact, it is much more efficient to focus on a subset of features. There are three main approaches to do this. The first involves the idea of distinguishing between features that are relative to target. The goal here is to determine a feature, or a set of features, that have the distinguishing characteristic. What it boils down to is: if there are two examples, it is important to single out the feature that can discern between them.

The next approach to feature selection is very similar to the last one discussed,

except it guarantees the example to be in the sample. It makes sure the probability is not zero for the feature in question.

The last idea on the topic was to recognize the importance of the rest of the features, not just the one in question. The relationship between features is what the determinant of the third approach. If for example, the feature in question is only important if the rest of the features are, then clearly it needs to remain. If however, some of the other features don't need to be there, it may alter the importance of the feature in question. Therefore, this approach encourages the re-evaluation of the remaining features with the removal of the irrelevant ones after every step.

So the next question is how does one eliminate the features that shouldn't be there? One option is to repeatedly remove features with the lowest weight, retrain the model and re-evaluate the performance. Hakenberg *et al.* discuss this method and call it Recursive Feature Evaluation (RFE) [17]. While building a feature set for a Support Vector Machine for gene name recognition, the goal was to eliminate all features and feature classes that decreased performance of the system while leading to a concise and informative model. To attain the best set of features that maximizes performance, 10% of the lowest weighing characteristics were repeatedly removed. The best results were attained at a feature set of 23, 000. However, as the goal was to also produce a concise set, the RFE was performed until only 150 features remained. These were said to be the most discriminating features.

CHAPTER 3

METHODOLOGY

The goal of this thesis is to establish and evaluate a classification model using a Support Vector Machine that could have the ability to predict micro RNA target binding sites. There are three main tasks in the process: define and build the necessary feature set, implement the Support Vector Machine, and evaluate the features. The following sections describe the steps necessary to achieve a working Support Vector Machine along with a useful presentation of the importance of the features independently and when implemented together.

3.1 Defining and Building a Feature Set

There are three features that are studied in this project. Each of the following three sections will give a bit of background on the characteristics and explain how the feature was built.

3.1.1 Expression Data

The expression data is the novel part of this experiment. Although gene expression as a topic is not anything new, an interesting phenomenon was observed by Dr. Chris Rock that soon became the main focus of this thesis. Gene expression itself is the process by which the information which comprises a gene can be realized. Some of the products that can be seen include proteins and RNA. Gene expression is a multi step

process which includes transcription, AKA “the molecular dogma” [8], the post-transcriptional modifications, and translation. In our case, we are studying micro RNA, and as mentioned before, miRNA is often categorized as a post-transcriptional negative regulator, thus, it follows that gene expression would be affected by the presence of miRNA. What is hypothesized in this work is that steady state levels of antisense transcripts to miRNA targets are generated through a miRNA- associated mechanism involving RNA- dependant RNA polymerases (RDRPs).

Antisense transcription is a recently discovered and validated phenomenon that is still not well understood. Relying on the general biological assumptions, DNA is transcribed into RNA and later translated into proteins. Each DNA part has 2 strands, the sense and antisense strands. The anti-sense strand of the DNA is transcribed into sense RNA by Watson: Crick base pairing rules that will eventually be translated into protein. The complimentary strand, the antisense, doesn't contain any protein making information. If the sense DNA strand is transcribed into RNA, it has the same sequence, aside from the replacement of Thymine with Uracil, as the anti-sense DNA strand, and is thus thought to be functionally useless. However, recent evidence supports the possibility of the anti-sense strand producing important effects through RNA interference, which involves the production of small RNAs similar in their biogenesis to miRNAs, except that they originate from the same locus that they silence. The function and origins of antisense transcription remain a mystery and are still being studied. This thesis and the hypothesis that caused the implementation of the work might be able to shed some light on the topic.

The interesting phenomenon witnessed by Dr. Chris Rock is being called the “ping-pong” effect. The “ping” is the antisense signal observed immediately upstream

(relative to the sense strand) of a miRNA target. The “pong” is the signal on the sense strand and is a known effect of endonucleolytic processing of miRNA targets by RISC. The sense strand signal has been studied more intensely and is better understood, while the antisense signal is the focus of this thesis. An increase in signal is seen downstream of the miRNA binding site on the sense strand and upstream (relative to the binding site on the sense strand) on the antisense strand. It is hypothesized that the source of antisense is in fact miRNA binding. Although it is known that miRNA plays a big role in gene regulation and developmental programs, perhaps antisense is actually responsible for a part of the effects. To explore the antisense transcripts as they relate to the miRNA targets, the expression data for the miRNA target genes must be extracted from the whole genome tiling micro-array datasets becoming more available with technical advances in the genomics field.

The building of the database was the most challenging part for the feature of gene expression because of the need for inter-disciplinary collaboration to extract the numbers for the analysis. Within the datasets that are available for the antisense transcription project, there are several files that contain all the information for the entire *Arabidopsis* genome. There are five main files that contain the data sets of the expressed sequences and two other files that are used for mapping whole genome expression data.

The five files of hybridization signals fall into two categories. The first consists of four files with sequences of 25 base-pairs resolution (bp) from RNA extracted from different tissues of the *Arabidopsis* plants. Each probe sequence was synthesized using photolithography by Affymetrix (www.affymetrix.com) [42]. The four files are organized by plant part: Flower, Root, Suspension Culture, and Leaf, or FL, RT, SC, and

LP respectively.

The last of the five files falls into a second category of data generated using a different micro-array platform. It is also a dataset with a series of sequences; however this file contains 36 base-pair long strands separated by 10 bp gaps (46 nucleotide resolution). It is called the T87 dataset and was synthesized by digital light processing by NibleGen Systems, Inc., (www.niblegen.com) [39]. This is a new technology that can be re-programmed at anytime and makes micro-array fabrication much more flexible. T87 is a suspension – culture sample similar to the SC sample.

The five files previously discussed are coordinated to the genome of *Arabidopsis* by the three data files. The Viktor file is a mapping file that is used with the expression data in the T87 file and has a 46 base pair resolution, genome wife on both strands. The Ecker file is used for the other four data files (the plant parts) and has 25 base pair probe resolution. These mapping files consist of lines of DNA sequence and show position, strand and chromosome number based on the genome sequence of *Arabidopsis*. The strand is either Watson or Crick, where the Watson strand is defined as the upper strand, and the Crick strand is the antiparallel (lower) DNA strand or partner strand that is complementary Watson strand.

Using Practical Extraction and Report Language software (Perl) with the dataset files, the expression levels of each gene on different parts of the plant were extracted using the third, Customsites file, containing the absolute location on the genome where the miRNA binds to the gene (literally the middle nucleotide of the miRNA). This information was extracted from the ASRP, Arabidopsis Small RNA Project website (asrp.cgrb.oregonstate.edu). The output file was constructed consisting of the gene name,

from which file (RT, SC, FL, LP, and T87) the number stemmed from, and the expression data on the sense and antisense strands that spanned 800 nucleotides upstream and downstream from the miRNA binding sites to validated miRNA target genes. The numbers were then normalized by summing all the row data for each strand over the target gene interval and dividing each datum by this value to give percent signal, before the Fast Fourier Transform was performed on all target genes to access average effects.

The control (null) set for the miRNA target genes was also created using paralog genes, which are evolutionarily related to the target gene and therefore biologically and structurally similar. However, the miRNA can not bind to the paralog because they have diverged over evolutionary time. Once the paralogs were determined for the genes by BLAST searches and manual inspection and assignment of “mock” miRNA coordinate binding sites, a file containing “binding site” locations was also built. Again, these are not where the miRNA bind, since it cannot bind to the paralog gene, but a plausible location based on conserved domains found in the paralog and target genes. The expression levels for the paralogs were extracted in the same manner as for the target genes. Based on the “binding site”, the numbers could be extracted from the 5 expression datasets and the corresponding genome annotation files. These values were also normalized, similarly to the target genes. Each row contained expression data for 800 base-pairs around the binding site, for both the sense and antisense strands were adjusted such that the sum of the values in a row was 1.

The goal of the analysis of the expression data is to identify high-confidence antisense transcription presence as a function of miRNA binding site, termed the “topology” of the sense and antisense transcription signals. If a relationship exists

between miRNA binding and antisense transcription, then the characteristic “ping-pong” signal could be searched genome wide, to predict candidate novel miRNA binding sites with a similar “ping- pong” topology.

The expression data were converted to Fourier Space using the Fast Fourier Transform algorithm [29].The Simulink toolbox from Matlab was used to apply the algorithm to the data and retrieve a pictorial representation of the ping-pong effect. After the presence of the effect was confirmed, the next step was to convert the dataset into a usable feature set for the Support Vector Machine.

Again, Perl was used to extract the necessary data and store them in a usable manner. Each of the expression levels at the locations ranging from negative 800 to positive 800 bp relative to the miRNA binding site in target genes was stored into its own feature number. This was done for every target gene on the sense and antisense strand. Therefore, feature one of each gene corresponded to the normalized expression level of that gene on the sense strand at 800 base pairs upstream of the binding site; feature 2 corresponded to the expression level at 775 bp upstream, and so on. The features continue to increase as the algorithm traversed the sense and the antisense strand, resulting in 130 individual features for the general feature of expression levels. The target genes were all assigned an SVM training value of positive one, and the paralogs with a negative one as training examples. Because a Support Vector Machine works as a binary classifier, there need to be both examples of what characteristics define a target gene, and which ones define it to not be a target gene.

3.1.2 Small RNA Counts

The second feature implemented for the SVM deals with small RNA counts with respect to the gene. Small RNAs are markers of RNA interference activity in cells and map often times to genes, in addition to paracentromeric and transposon repeats in genomes [21, 22, 11, 31, 26]. This observation is another type of indirect evidence for antisense transcription because small RNAs are produced from double stranded RNAs. The small RNAs indicate transitivity in part due to antisense transcription, and therefore can be tested as another feature for miRNA target site prediction. Although it is not directly related to the miRNA target sites, if following the hypothesis that antisense transcription and miRNA target site binding are related, the abundance of small RNA should be a valid indicator. Dr. Chris Rock performed a statistical analysis of the abundance of small RNAs mapping to miRNA target genes compared to paralog genes and found there were significantly more small RNAs associated with targets than non-targets [37]. In this thesis, the small RNA feature actually represents the number of expressed distinct signatures obtained from different tissues. In other words, the numbers seen here represent the different signatures that matched the specified gene, not the total counts which would include independent isolates of the same small RNA. Distinct signatures are therefore a conservative, quantitative method.

Building the small RNA feature was done with the use of MPSS data obtained through a web portal [25].The list of potential miRNA target genes was processed by a bulk query and the output included distinct signatures from several libraries made from different tissues. The libraries that were used for this work were the inflorescence tissues and the seedlings. According to the Library Submitter, Cheng Lu, the plants were grown

in a chamber, in soil, with 16 hours of light for 5 weeks. Immature inflorescence was harvested about two hours after dark and the tissue studied was inflorescence meristem and early stage floral buds. The TRIzol reagent was used to isolate the RNA.

The second library contained information about the seedlings. Here, Cheng Lu explains that the seedlings were grown on media with 1 xMS salts, 1% sucrose, and 2 week old seedlings were sprayed with water. After 2 hours, seedlings were collected and total RNA was isolated using the TRIzol reagent again.

The data analysis was done with Perl, extracting the necessary information and producing the sum of the data from both libraries. The small RNA feature was incorporated with the data from the expression levels by adding it as the next feature for the corresponding gene. Therefore, the small RNA counts became feature number 130 (the first 130 fit between locations 0 and 129). A large number of the miRNA target genes did not have small RNA counts associated with them, although this does not dismiss the possibility of the gene being a target. If there was no small RNA count associated with the gene, feature 130 was left blank and the Support vector Machine interpreted it as if it was filled in with a zero.

3.1.3 Thermodynamic Feature

The established method of quantifying miRNA complementarity to the target gene is accounted for within this feature. Although plants have an extraordinarily high complementarity between the miRNA and the gene, it is still difficult to define a perfect threshold that would not only allow for some error, but be stringent enough to not

increase the number of false positives. Although base-pairing would be quite effective within Arabidopsis, it is much more effective in the long run to implement an algorithm using a feature that would not only account for complementarity but produce results with a "grade" of how well the teams of features go together.

The goal should not be to identify the greatest number of possible matches, but the best matches. Because complementarity-based algorithms find all possible matches between the miRNA and the entire genome, it proves to be not economical to add more potential targets to the list. Since every target has to be confirmed by wet lab experiments, this could prove to be quite costly. The evaluator for the feature of miRNA and the RNA interaction that seems to be more useful is that of free energy. Instead of finding every sequence that seems to match the miRNA based on a certain threshold, resulting in countless possibilities, it is better to find a stable version of binding between miRNA and the corresponding target gene.

To attain the free energy values, several steps had to be taken. Target gene sequences were extracted and connected by a series of seven Uridines to the matching miRNA sequence. This final string was analyzed by the Quickfold algorithm, an RNAfold variation on the DINAMelt server [24]. Multiple sequences could be bulk processed and the most negative value was used for the particular sequences. The most negative value corresponds to the most stable combination of the miRNA target gene and the miRNA itself.

Next, the string combinations were perfected. Since the complementarity of the miRNA and the RNA isn't perfect, those errors were corrected, by yet another Perl script. The bases were made to match up perfectly and the new strings were again analyzed by

the DINAMelt server. The resulting numbers were a bit higher than those of the original energy evaluation. Because the sequences would bind perfectly, with no mismatches or errors in complementarity, it makes sense that the binding energy would be more stable. The energy received from this analysis became the denominator to our function. The numerator was the "imperfect" energy, the target to miRNA match as it was originally. The result, the ratio of original free energy to the free energy if the miRNA had perfect complementarity to the gene (percent maximum free energy), became the final feature for the Support Vector Machine. Feature number 131 was filled in with the decimal between zero and one for each of the corresponding genes that already contained expression and small RNA count data.

A similar process was done for the paralogs. However, it became a bit more tedious and more work. Although the miRNA sequence remained the same, the paralog sequence had to be manually extracted from the gene. This added several steps to the process because the entire sequence had to be found, and then the "binding site" had to be manually located. However, after the sequences for the paralog gene and the miRNA were retrieved, the rest of the process remained the same. The same algorithms were used to connect the sequences and run on the DINAMelt server.

3.2 Support Vector Machine

As discussed before, a Support Vector Machine (SVM) is a tool to classify certain objects. In this case, the SVM was used to classify miRNA target genes. After a database was built with the features consisting of expression data, small RNA counts and energy

values for each of the possible target genes and the paralogs, all that was left was to run the algorithm. Ten-fold cross validation analysis was performed. The complete set was split into ten parts and analyzed repeatedly with nine of the ten sets being used as a learning set and the last of the ten as a testing set. The process was repeated 10 times such that each of the 10 sets could a test set. The Support Vector Machine tool used for this process was a Matlab SVM toolbox [6]. It is an implementation of Vapnik's SVM, in this case for pattern recognition, in C.

As mentioned earlier in the chapter, the target genes all had a label of positive one, while the paralogs had a label of negative one. After the label the feature values followed. The SVM is able to develop a model based on the label and the features that should be able to discern between a plausible target gene and a gene that does not follow the pattern to be a target gene. Based on this idea, the dataset created was analyzed and a model was built for each of the 10 combinations of sets.

CHAPTER 4

RESULTS

This chapter discusses the results of the analysis of the features of the Support Vector Machine. The evaluation of the features is emphasized and possible explanations for the results are provided.

4.1 Fast Fourier Transform

Based on the idea that there exists a correlation between antisense signal and miRNA binding, the first goal of the research is to document the "ping-pong" effect, a phenomenon that could be used to later predict miRNA target sites. As previously discussed, Dr. Chris Rock found an indication of increased signal for antisense transcription within 200 base pairs of the miRNA binding site on possible miRNA target genes. The data was analyzed with Fourier Analysis to more clearly show and support the presence of the signal.

The image in Figure 4.1 is a plot of expression levels for each of the possible target genes. The expression levels are centered about the binding site, at position 0 with 400 base pairs upstream and downstream. The red data represents the expression levels of the sense strands, while the blue is for the antisense strand data. There is no useful information represented in this figure. No pattern can be seen in the expression levels and it appears to be a noisy signal.

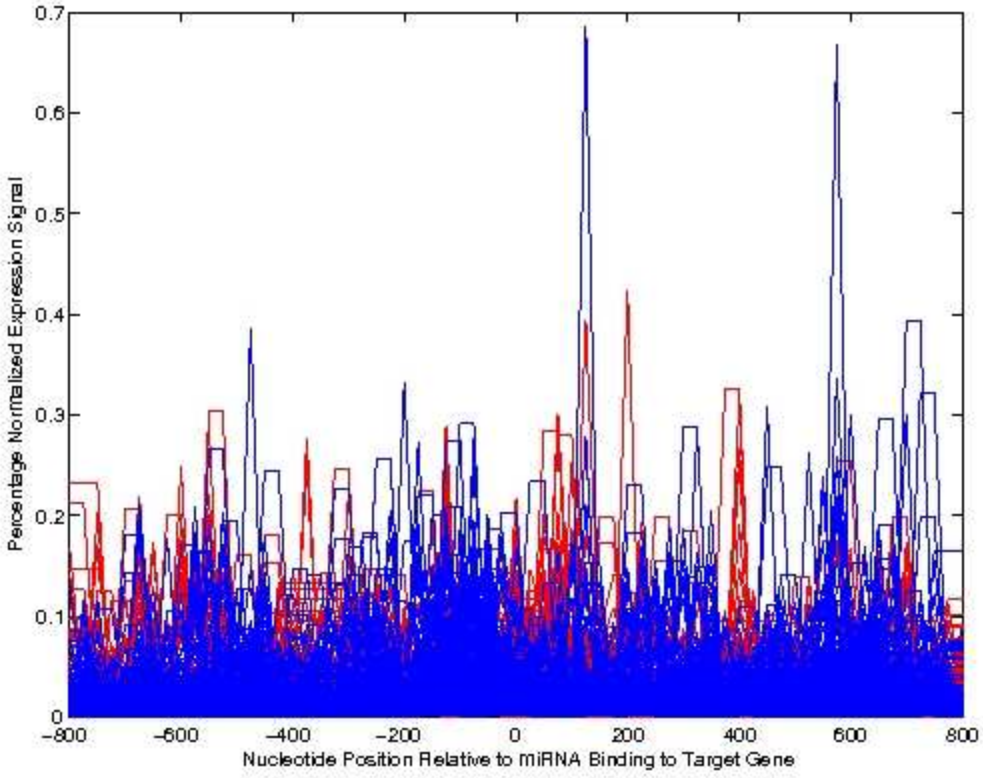

Figure 4.1: Topology of Expression Levels Before Fast Fourier Transformation. The blue lines correspond to the expression levels on the antisense strand of each gene, while the red lines represent the expression levels on the sense strand for the bonafide miRNA target genes.

Figure 4.2 illustrates the "ping- pong" effect to a degree. Since the goal of the Fast Fourier Transform is to drown the "noise", so to speak, it is evident from Figure 4.1 to Figure 4.2 that the random peaks were minimized and the final result shows corresponding peaks in Figure 4.2 around the "ground zero", zero on the x-axis. Once the numbers were transformed into Fourier space, the increased expression signal became clear. Since each of the lines corresponds to the sense and antisense data, the graph demonstrates the matching, but transposed peaks that define the "ping-pong" phenomenon.

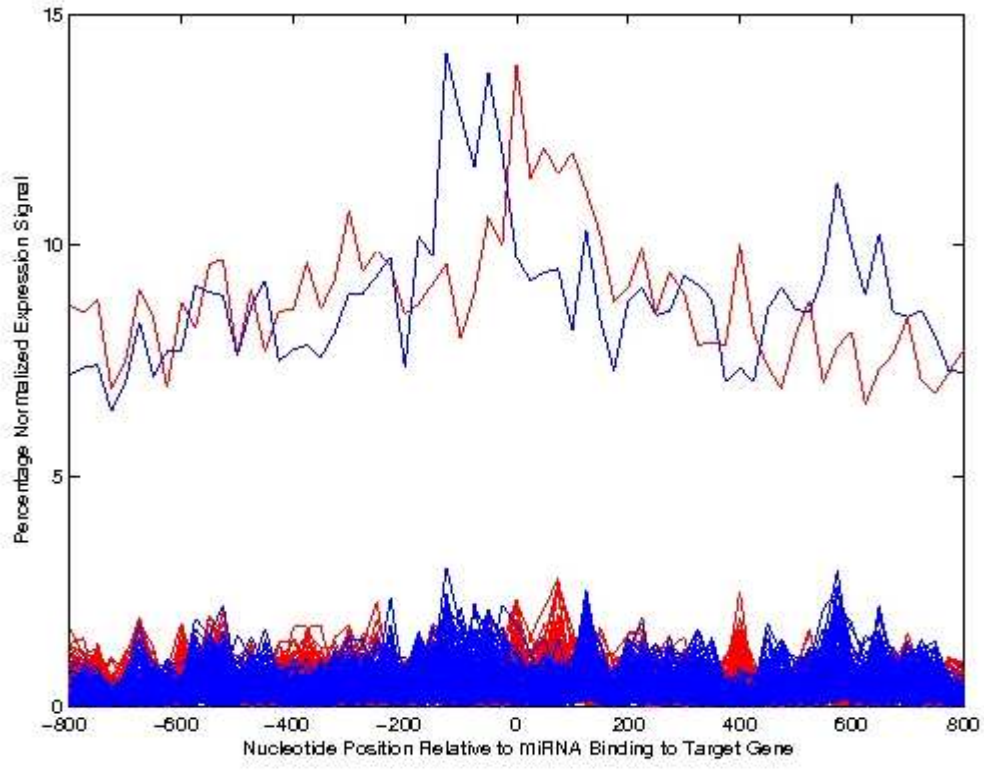

Figure 4.2: Expression Levels After Fast Fourier Transformation. The red line corresponds to the sense strand of the target genes and the blue to the antisense strand. The blue and red line seen at the top of the figure are the sum of the percent normalized expression signal after the Fast Fourier function had been applied to each value.

The miRNA target data were compared to that of the non-target paralogs. As stated before, these are similarly structured genes that would in theory have similar functionality and therefore serve as suitable biological controls. The FFT analysis shows these genes, but lacks the expression level pattern seen in the miRNA targets. This set of data was used as a control set for the research. The data was similarly extracted from the previously discussed files and normalized in exactly the same way. The results of the Fast Fourier Transform support the hypothesis that miRNA targets have a novel antisense activity while paralogs lack the corresponding peaks around the miRNA binding site

(zero position).

The image before performing the FFT on the paralogs looked very much like Figure 4.1 when using the miRNA target dataset. However, after the FFT, Figure 4.3 does not seem to demonstrate any extra information, even with the noise decreased. There are no corresponding peaks around the hypothetical binding site and the sense and antisense data seem to not show anything interesting.

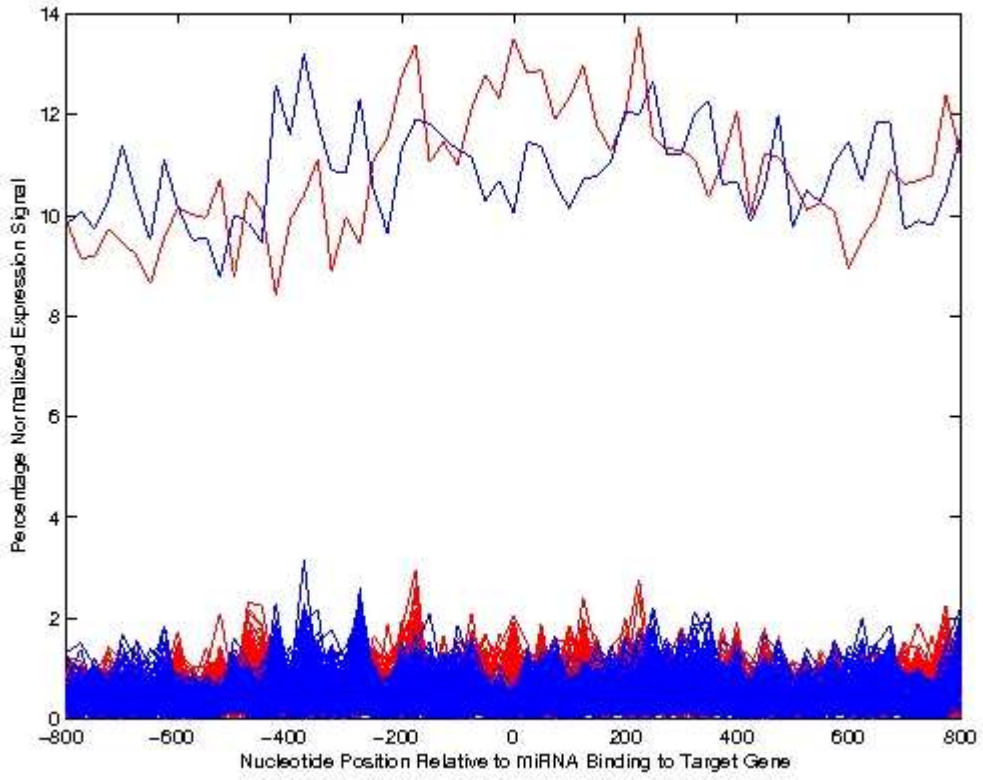

Figure 4.3: Expression Levels of the Paralog Genes. The red line corresponds to the sense strand of the target genes and the blue to the antisense strand. The blue and red line seen at the top of the figure are the sum of the percent normalized expression signal after the Fast Fourier function had been applied to each value.

4.2 SVM Evaluation

The FFT results are evidence to support the existence of the "ping-pong" effect. Based on this finding this phenomenon was then added as the key feature into the feature set for the Support Vector Machine to explore the importance of the heightened antisense signal. The complete data, with the target genes and the paralogs was put into a format that could be run with the SVM toolbox. The data was then analyzed through ten fold cross validation. The various combinations of features were analyzed to evaluate the importance of each in identifying correctly the miRNA target genes.

The notion of the Gold Standard can be incorporated into the analysis. In machine learning such as the Support Vector Machine, where the goal is to classify samples, the Gold Standard refers to a set of data that can be used to validate the results. In the research done here, the data used, is in fact a set that could be used to evaluate how well the Support Vector Machine can classify each of the genes as a possible target or not. Because the data consists of labeled, confirmed target genes and the negative control paralogs, there is now a basis for positives and negatives. The target genes are positives and the paralogs are negatives. Implementing the SVM produces an output file consisting of the predictions the machine made. This file is composed of data corresponding to each of the genes, with a confidence score of which class it belongs in. The options for classification are +1 or -1, where, as stated before, the target genes correspond to the positive number and the paralogs to the negative. In the output file, if the prediction score is 0.2567, the gene is classified as a +1, or target gene, if it is a -0.2567, it is classified as a paralog. However, if the gene was a 2.567, it would still be a target gene, but the machine predicted that with a much higher confidence level. Because the classification is

done on a range of confidence levels, there are going to be errors made within the output. Thus, using the Gold Standard, the evaluation can take place.

4.2.1 Accuracy

After the support vector machine was run ten times for each feature, with the testing being done on one of the previously made sets, the accuracy and the standard deviation were calculated. Table 4.1 demonstrates this statistic, one of the tests that could be performed using the Golden Standard.

Table 4.1: Average Accuracy and Standard Deviation. The averages and standard deviations of the ten sets were calculated for various combinations of features.

Combination	Average	Standard Deviation
Expression Levels, Small RNA Counts, Energy	0.8555	0.0344
Small RNA Counts, Energy	0.8516	0.0305
Expression Levels, Energy	0.8547	0.0336
Expression Levels, Small RNA Counts	0.618	0.0163
Small RNA Counts	0.5999	0.0016
Energy	0.8477	0.0346
Expression Levels	0.5779	0.0331

The accuracy of the Support Vector Machine can be evaluated by this equation:

$$\text{accuracy} = \frac{\text{number of True Positives} + \text{number of True Negatives}}{\text{numbers of True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}}$$

True positives are the numbers represented as those that were predicted to be positive, or

the target gene, and actually are labeled, based on the Gold Standard as the target gene. True negatives are similarly the genes that are labeled and predicted to be negatives, or the paralogs. The rest are errors, false positives are actually negatives, but predicted to be positives, and the false negatives are positives that were predicted to be negatives.

The results can be interpreted to represent the importance of each of the features. When looking at energy alone as the prediction mechanism for miRNA binding sites, the success rate is 84.7 %. This is a feature that has already been confirmed to be very important in small RNA binding site prediction. The high performance is completely logical and expected. However, when paired with the small RNA counts, the accuracy of the prediction using the Support Vector Machine increases to 85.1%. The increase in accuracy can be interpreted to mean that small RNA count improves performance of the Support Vector Machine, and thus it is also an important feature. Similar results can be seen with the other features as well, including the addition of the ping- pong effect. When the expression data are added to the feature set, the test shows an even greater improvement in performance than when the small RNA counts were added. Again, this is seen as support for the newly discovered phenomena being indicators of miRNA binding. Independently, the accuracy of the small RNA counts and the expression data are approaching random. However, it is important to note that in combination with the often used feature of energy, performance is improved. Even still, when all three features are tested together, the accuracy of the Support Vector Machine improves even more. When combined, it becomes clear that the features are important to the prediction of miRNA target sites.

4.2.2 Specificity, Sensitivity, and Precision

Further evaluations of the performance of the Support Vector Machine were done. Also based on the Gold Standard, Table 4.2 displays more of the statistical tests that are possibly more useful in evaluation of the features.

Table 4.2 Average Sensitivity, Precision and Specificity. Values were calculated of the ten sets for various combinations of features.

Combination	Sensitivity	Precision	Specificity
Expression Levels, Small RNA Counts, Energy	0.7393	0.924	0.9498
Small RNA Counts, Energy	0.741	0.9122	0.9416
Expression Levels, Energy	0.7393	0.9219	0.9484
Expression Levels, Small RNA Counts	0.2198	0.7742	0.946
Small RNA Counts	0.1863	0.7525	0.938
Energy	0.7462	0.8981	0.9297
Expression Levels	0.068	0.9667	0.9969

Specificity refers to how well a classification test can identify the negative cases. In other words, this evaluation represents the probability that the classifier will classify the gene as -1 if the gene is a paralog. A high percentage is important if it's needed to know that the genes will not be a target, but a paralog.

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

The sensitivity, on the other hand is the evaluation of the test to predict the targets, or the positives. In other words, it is the probability that the SVM will predict a

gene to be a target correctly.

$$\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}.$$

The Positive Predictive Value, or the precision, addresses the evaluation of the machine. The number will represent the probability that if the SVM stated the gene to be a target, how likely it really is to be a target gene. This test is the reverse of the previous two. The sensitivity and specificity test the machine in the respect of if the actual label is know, how likely is it to identify it correctly. Whereas the PPV evaluates the machine by if the machine states the gene is a target, how likely is it to really be a target.

$$PPV = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Positives}}$$

The three tests are all related, just in a different order. The sensitivity and the precision test both look for accuracy of identifying a true positive, but sensitivity evaluates it against false negative and the precision evaluates it against false positives.

All three tests, with respect to the data used in the thesis, display, that no matter what the combination of features is, the machine can identify the true negatives well when the gene really is a negative (paralog). Also, when energy is involved in combination with either the small RNA count or expression level, or both, and the machine states the gene is a target, it has a good probability of really being a target.

The sensitivity of the Support Vector Machine, however, is not as high as the other two values. If the label of the gene states that it is a target gene, the machine is more likely to actually predict it as one if more features are used, as expected. Small RNA counts improve sensitivity levels of the experiments. Although energy plays the biggest role in the performance, small RNA counts give a slightly better value in combination with energy than the expression levels do.

When broken down by combinations, a compelling statistic can be seen. It appears that the SVM is better at categorizing a gene as a paralog than a true target gene. This could possibly be a good thing, perhaps it means there will be less genes declared targets that wetlab results would prove untrue. In general, it appears that expression levels and small RNA counts play a role in increasing sensitivity. However, the energy feature is important also because in the combination of counts and expression levels, with no energy, the false negatives give a greater number. It has already been established though that the energy feature is a very useful characteristic, based on complementarity for developing a prediction mechanism. What is not established, but there is evidence for from this analysis, is that the new features, antisense expression signal and small RNA counts, contribute greatly and complement the energy feature in building a useful predictive model.

CHAPTER 5

FURTHER WORK

By this time, all the work that had been done could be used as support for the importance of the newly identified qualities seen in target genes. It is evident that the "ping-pong" effect, the heightened signal for antisense transcription upstream of miRNA binding sites, was indeed real and a contributing part to the prediction mechanism. However, using only our data was potentially limiting and new test data were the logical next step

5.1 Building the Database

A new database is warranted to test the Support Vector Machine that has been built. Such a resource was found in the work of Adai, *et al.* [1]. Their work has contributed to building a website centered on predicted miRNA and precursor candidates for the *Arabidopsis* genome. An algorithm they developed, called findMiRNA, is a powerful tool that searches for characteristic divergence patterns in candidate miRNA gene families and for conserved target sites within related transcripts from other species than *Arabidopsis*.

The test dataset that was built resulted from an output of findMiRNA. The process first uses a rigid complementarity threshold to identify initial miRNA candidates. Then the algorithm "analyzes the candidates for adjacent sequence complementarity that would enable stem-loop formation in an RNA molecule consistent with the known structures of

miRNA precursors" [1]. A set of 1599 clusters was generated from a scan of 5701 transcripts. The resulting group consisted of a ranked set of families based on sequence conservation of the miRNA and the target gene. The final set contained potential target genes for the miRNA with the highest ranks at the top. A Perl script was then used to extract the necessary information to build the dataset necessary for the Support Vector Machine. Because it has already been shown that small RNA counts do not make a large impact on the accuracy and other evaluations of the machine when energy and the expression levels are involved, this feature was not included in the work discussed in this section.

5.1.1 Expression Levels

The Adai database was not just a useful set of genes, but also a useful starting point for extracting the needed information to fill in the expression data for each of those genes. The database listed the gene and the relative location of the binding site for the miRNA. Essentially what this means, is the number given was where the start of the miRNA sequence lines up with the particular location on that gene. Therefore, if the relative location is 15, that means, 15 nucleotides from the start of the gene, the complementary sequence to the miRNA begins.

This however, was not the form of the location that was needed. An absolute location was what was necessary to extract the expression data using our existing Perl scripts. A position on the genome was necessary, so if the gene starts 300 nucleotides from the start of the genome, the previously discussed theoretical gene that is 15

nucleotides from the start of the gene, would have the absolute location of 315.

A Perl script was developed to extract gene names and the relative starting and ending points from the Adai database. Next a Python script was developed to find the absolute location within the TAIR7 release of the gene features of the Arabidopsis genome. This was done by a means of hashing and looping to keep track of the gene, what strand the gene is on (the sense or the antisense strand) and the absolute location of it. Next, the file that has been extracted from the Adai database was looped through to calculate which exon the relative location mapped to and what the corresponding absolute location was. The start of the location, however, was not the number we were looking for; we also had to add 10 to it. Since the binding site, in our database has been "ground zero", and the sequence lengths of our miRNA and gene snippets have been 21 base pairs long, the 10 had to be added to get to the middle of the strands. Then, the data would match our original database with 10 base pairs up and down stream of "ground zero".

Once the absolute location had been determined, it was a matter of extracting the expression data from the files that had been already built as discussed in Chapter 3. Overlaps form during the compilation of expression data due to the process of retrieving the expression numbers, thus some of the cells at certain positions within the 800 range up and downstream of ground zero result in blank spaces. These rows of expression data were removed from the final dataset to prevent inaccurate numbers from being used in the Support Vector Machine.

5.1.2 Energy

The next feature, the thermodynamic energy of binding between the miRNA and the target gene was the next step in building the dataset. Again, the Adai database was the starting point. The target gene, the matching miRNA and the corresponding sequences of the hairpin were extracted. With the script from the earlier work, the miRNA sequence was transposed to be in reverse order and connected to the target gene sequence with a series of seven Uridines to form the head of the pin. The resulting string of the sequences from the target gene and the miRNA was the input for the DINAMelt server which predicts melting profiles for nucleic acids. The algorithm was used for fast folding of the sequences to obtain the binding energies [16]. Under the RNA conditions of version 3.2, the most stable energy for each of the sequences was retrieved. This resulted in numbers that correspond to the energies that were retrieved for the original dataset. Next, the matches between the gene and miRNA were made perfect, as if it would be a 100% base pair wise match between the two structures. Again, the sequences were analyzed with the DINAMelt server to get a second energy. The two were divided to find a percent maximum free energy. This ratio was used as the second feature for the support vector machine for the new dataset.

5.2 Support Vector Machine

The two features were combined by another Perl script and put into the correct format to be used by the SVM toolbox in Matlab. As in the original dataset, discussed in Chapters 3 and 4, the result was a series of 130 features for the expression levels.

However, since the small RNA counts were dismissed, feature 130 (the 131st feature, however, numbered 130 due to the feature indexing beginning at zero) was set to zero. Feature 131, as before was filled in with energy ratio values and the input for the SVM was complete.

The dataset discussed in the previous chapters became the training set for the analysis of the dataset formed from Adai's work. Instead of performing the split of the original data, it remained in one group, while the resulting dataset from the Adai paper became the testing data for the Support Vector Machine.

5.3 Results

The Adai dataset was analyzed using the Fast Fourier Transform to verify the presence of the "ping-pong" effect. Figure 5.1 demonstrates the noisy signal.

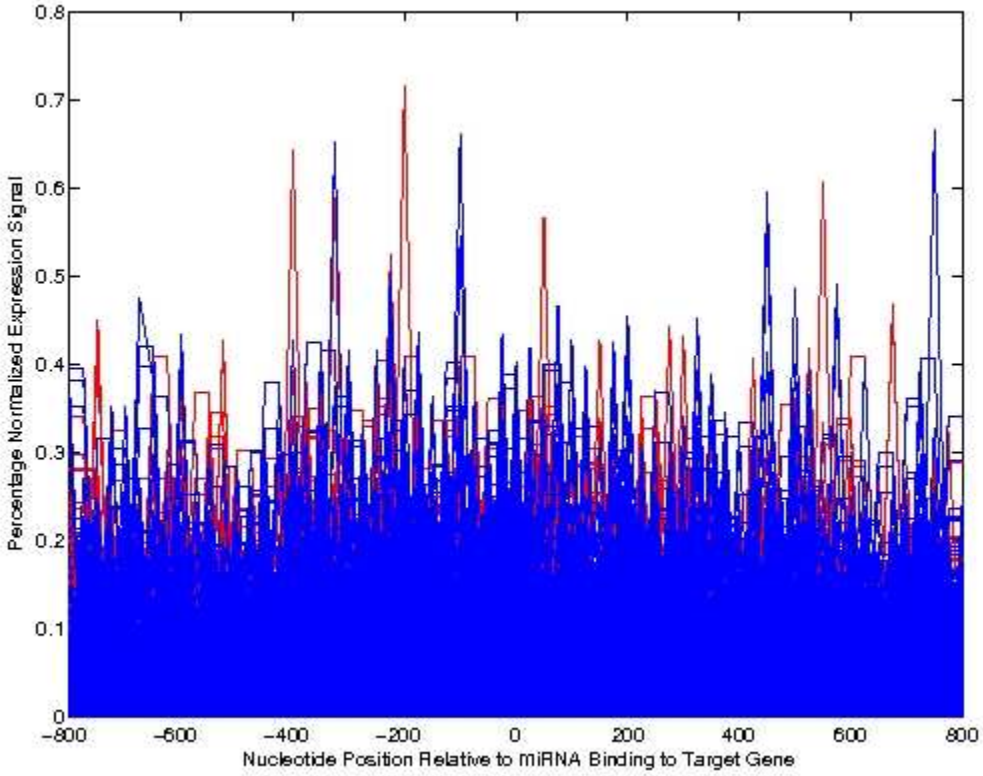

Figure 5.1 Expression Levels Before Fast Fourier Transform. Expression levels of genes for the Adai based dataset are represented. The blue lines correspond to the antisense strands and the red lines correspond to the sense strands.

Similar to the original work, the graphical representation of the expression levels is very noisy and no real value can be placed on the data. However, after the Fast Fourier Transform was applied to the expression levels of the genes in this particular dataset, a clearer signal was witnessed as shown in Figure 5.2.

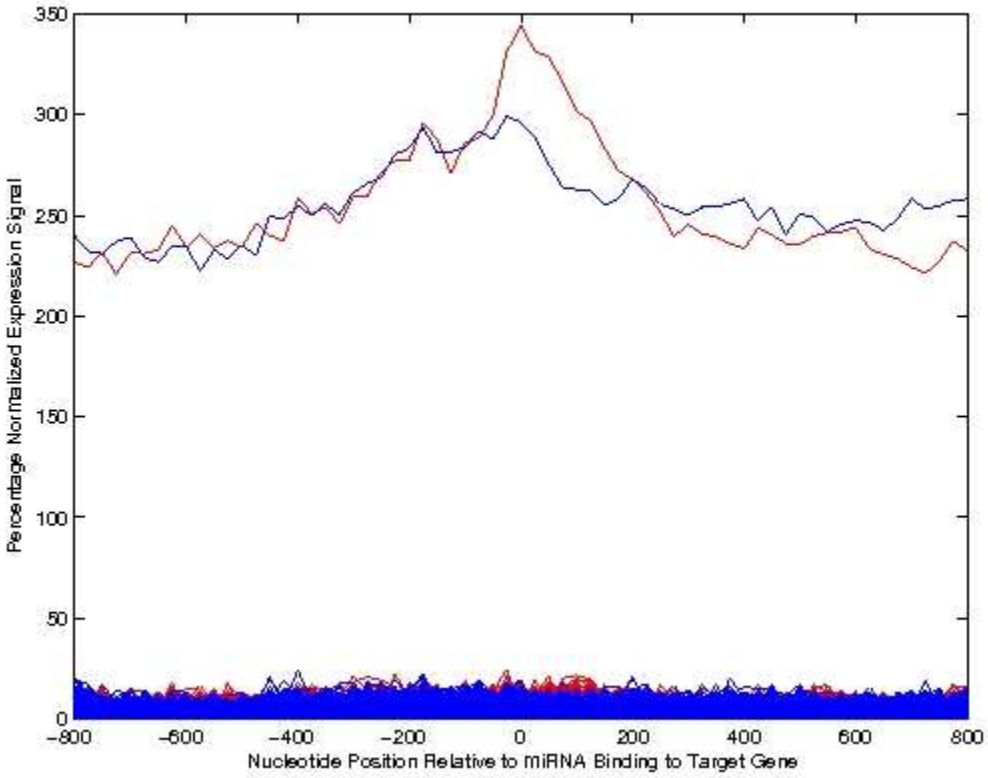

Figure 5.2 Expression Levels After Fast Fourier Transform. Expression levels of genes for the Adai based dataset are represented. The blue lines correspond to the antisense strands and the red lines correspond to the sense strands. The upper lines correspond to the summation of the signals in a specific location.

After the transform was applied, the noise in the signal was reduced and the upper strands seen in Figure 5.2 represent the clear signal. The genes in the dataset being analyzed are not verified, therefore the “ping-pong” effect cannot be clearly seen in the figure. However, it is still interesting to see peaks in the 200 base pair region both up and downstream of the binding site. From this evidence, it could be extracted that some of the genes could, in fact be true miRNA target genes and therefore display the corresponding peaks in the region of interest.

In the Support Vector Machine analysis, 1275 examples were used as the training set to learn the prediction model that resulted in 649 support vectors that would be used on the testing set. The testing set consisted of gene expression levels from 5 sources, the flower, root, suspension culture, and leaf files, consisting of 25-mers, along with the expression levels from another suspension culture sample that consisted of 36-mers. As previously done, rows with any missing values for a specific location were thrown out and the resulting testing data consisted of 16606 entries.

While formatting the test data for the SVM, an initial category prediction had to be provided. The analysis of the accuracy of prediction would be easiest if all the entries had the same prediction, so it was decided that all the genes were to be predicted to be miRNA targets, or the classification of positive one. The opposite classification would also work; negative one would correspond to the prediction that the gene is not a valid target gene for miRNA. However, because the Adai database used consists of genes that were declared as potential targets, I chose to label with a positive one.

The statistics from this analysis can be seen in Table 5.1. The values corresponding to each of the statistics evaluate the performance of the Support Vector Machine and demonstrate the extent of the agreement between the predicted values of the SVM and Adai's work.

Table 5.1 Accuracy, Sensitivity, Specificity, and Precision of the Support Vector Machine performance on the Adai based dataset.

Statistic	**Value**
Accuracy	0.9807
Sensitivity	0.9807
Specificity	NaN
Precision	1.0

Specificity does not contain a valid numerical value, meaning that a division by zero had occurred. When analyzing the breakdown of predicted values based on the dataset, this makes sense. The numbers associated with the predicted values of the Support Vector Machine can be seen in Table 5.2.

Table 5.2 Distribution of predicted values versus the actual label

		Condition (as labeled)	
		True	**False**
Prediction	**Positive**	True Positive = 16286	False Positive =0
	Negative	False Negative =320	True Negative =0

The false positives and true negatives are lacking numbers. Because all the test data were categorized as target genes, or positives, there are no true negatives that could

be calculated. Because of the same reasoning, there are no false positives. To be a false positive, the gene had to be originally categorized as a paralog, or a negative. Since there are no genes categorized as such, there cannot be a false positive. However, as Table 5.2 shows, of the 16606 predicted target genes according to Adai, 320 of them were predicted to be paralogs by the SVM, producing the 2% discrepancy in agreement.

The research done here, using the idea of “ping-pong” effect and thermodynamic energy of binding supports the results of the Adai *et al.* research which used different predictors. These results suggest that not only is the set of features discussed in this thesis are valid indicators for miRNA targets, but also there is now more support for the list of genes produced by Adai's research to be potential miRNA target candidates.

The research done in this thesis and the feature set that was developed seems to provide more stringency when predicting miRNA target genes. The accuracy, however, is still relatively high because the main feature implemented in the work for this thesis and Adai's research is shared. Thermodynamic energy of binding, as mentioned before, is the most commonly used indicator for miRNA binding; thus, the agreement between the two sets of data is inevitable. However, because the agreement isn't 100%, this means that according to the predicted values found by the Support Vector Machine, some of the genes that Adai and his collaborators predicted to be target genes are not target genes.

CHAPTER 6

CONCLUSION

This thesis has explored the use of the Support Vector Machine with certain features to predict miRNA target gene binding sites. The expression data, small RNA counts, and thermodynamic energy of binding were used as features within the Support Vector Machine to build a model for prediction. Using the original dataset as the Gold Standard, the machine was tested based on accuracy, precision, sensitive and specificity. The machine was further tested with a dataset of candidate miRNA targets from another lab.

The "ping-pong" effect was the new feature being incorporated into a set of previously proven mechanisms for target gene prediction. The phenomenon discussed displayed increased signal 200 base pairs upstream and downstream on the sense and antisense strand, respectively, of the miRNA binding site. This correspondence was put into numerical terms by extracting expression data 800 base pairs upstream and downstream of the binding site, on both the sense and antisense strand of the target genes. This resulted in 130 values associated with each of the genes in question.

The next feature, value number 131, consisted of small RNA counts. The data were extracted from the MPSS website and the small RNA counts seen for a particular gene were combined from the flower and the seedling samples.

The last feature incorporated addressed the issue of complementarity to some extent. Because plants possess the quality of near perfect complementarity between miRNA and the target gene, the feature was important to add to the prediction

mechanism. The most effective way of compensating for this need was to evaluate it on the thermodynamic level. The most stable combinations of miRNA and the target gene normalized to percent maximum free energy were the values that were added as feature number 132.

The dataset was built using Perl scripts to not only extract the needed information but to also put it in the correct format for the SVM toolbox. Ten-fold cross validation was performed on the dataset. The results were consistent with the working hypothesis and demonstrated that all three features play an important role in the prediction abilities of the machine. Each of the features when tested alone displayed a near random distribution, meaning that the prediction was no better than chance. However, as the features were combined, it became evident that expression levels and the thermodynamic energy of binding play the most important role in accurate and useful criteria for prediction of whether a gene was a miRNA target.

The next step of the process was testing the new mechanism on outside data. Using the Adai database, the expression levels and the energies were found for the genes the lab had listed as those of interest for novel candidate miRNA targets. The dataset that was originally used for testing and learning was used as a whole for the learning process in the test. After the model was built, the Adai based dataset was used as the testing set. The results were not as expected, not all of the genes on the list were accurately identified as potential miRNA target genes, but most were agreed to be potential targets. Based on these results, the Adai findings were substantiated using a different set of criteria (visual expression levels), which provide additional evidence for the validity of the features that were developed for this thesis and the Support Vector Machine.

There are certain limitations within this work, however. Each of the features, although clearly important and useful could be improved upon. The improvements would increase the accuracy and usefulness of the Support Vector Machine without decreasing its usability and efficiency. Addressing the first feature, the expression levels around the binding site, the improvement that could be made is more precise absolute locations. Originally the binding sites were calculated manually, so there is some room for error, although it is compensated for by the resolution for expression level signals being 25 base pairs. Improvement in micro-array technology that makes the resolution higher would make the expression values more precise. Automating the process of finding the miRNA binding site coordinates within the genome would reduce errors that could add to the overall error of expression values for each of the intervals.

Addressing the small RNA counts, the values should be updated. Because there are new data forthcoming on small RNA counts associated with each of the *Arabidopsis* and other genomes, the MPSS website is constantly updated. The dataset that was built should be updated to the changes that already exist. It might also be useful to incorporate other samples as values for the overall feature of small RNA counts. Right now, the sum of values for the flower and the seedling account for just one feature associated with the gene. Adding values from other tissues of the plant, and mutant genotypes might improve the Support Vector Machine.

The thermodynamic energy is perhaps the feature needing the least work. It is already shown to be a valuable addition to the prediction mechanism, improving the precision and accuracy to a great extent. However, there are some missing values for the target genes and the paralogs. If the gaps are filled in, it is expected that an improvement

will be easily seen in performance.

The next steps in research should not only address these improvements but continue to further testing and database expansion. Newly discovered miRNAs can be used to build a dataset with target gene expression data, small RNA counts and energies that could be analyzed by the support vector machine to find novel information. Because they have not been heavily implemented, perhaps the Support Vector Machine will be a new tool that could discover viable target gene sites for the miRNAs.

Another topic of exploring the importance of antisense signal in miRNA biology is using miRNA genes themselves as candidate target genes. There is evidence that leads to the possibility of intra-molecular binding between primiRNA and mature miRNA. The Support Vector Machine could shed some light on the topic.

The last and perhaps the most impacting expansion of work should be analysis of a different genome. Now that evidence supports the effectiveness of the Support Vector Machine that was built for *Arabidopsis thaliana*, applying it to the rice and other genomes should be the next step. If the results are promising, studying genomes even further away from these might show results that are compelling and revolutionary.

BIBLIOGRAPHY

[1] A. Adai, C. Johnson, S. Mlotshwa, S. Archer-Evans, V. Manocha, V. Vance, and V. Sundaresan. Computational prediction of miRNAs in Arabidopsis *Thaliana. Genome Research,* 15:78-91, 2005.

[2] E. Bernstein, A. Caudy, S. Hammond, and G. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature,* 409(6818):363-6, 2001.

[3] A. Blum and P. Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence,* 97(1): 245-271, 1997.

[4] K. Bohmert, I. Camus, C. Bellini, D. Bouchez, M. Caboche, C. Benning. AGO1 defines a novel locus of Arabidopsis controlling leaf development. *Embo Journal.* 17:170, 1998.

[5] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2(2): 1384-5810,1998.

[6] G. Cawley. Support Vector Machine Toolbox, v0.55beta. *University of East Anglia, School of Information Systems,* 2000.

[7] Francis Crick. On Protein Synthesis. In *Symp. Soc. Exp. Biol. XII,* pages 139-163, 1958.

[8] Francis Crick. *What Mad Pursuit (Alfred P. Sloan Foundation series),* 1988.

[9] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. Marks. Micro RNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

[10] D. Errampalli, D. Patton, L. Castle, L. Mickelson, K. Hansen, J. Schnall, K. Feldmann, D. Meinke. Embryonic Lethals and T-DNA Insertional Mutagenesis in Arabidopsis. *Plant Cell*. 3: 149, 1991.

[11] N. Fahlgren, M. Howell, K. Kasschau, E. Chapman, C. Sullivan, J. Cumbie, S. Givan, T. Law, S. Grant, J. Dangl, J. Carrington. High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of *MIRNA* genes. *PLoS ONE*. Feb 14;2(2):e219, 2007.

[12] R. Gregory, T. Chendrimada, N. Cooch, and R. Shiekhattar. (2005). Human RISC couples micro RNA biogenesis and posttranscriptional gene silencing. *Cell*, 123(4):631-40, 2005.

[13] D. Grierson, R. Fray, A. Hamilton, C. Smith, C. Watson. Does co-suppression of sense genes in transgenic plants involve antisense RNA? *Trends Biotechnol* 9: 122-123, 1991.

[14] S. Griffiths-Jones, R. Grocock, S. van Dongen, A. Bateman, and A. Enright. MiRBase: Micro RNA sequences, targets and gene nomenclature. *NAR,* (34): 140-144, 2006.

[15] S. Griffiths-Jones. The micro RNA Registry. *NAR* (32):109-111, 2004.

[16] A. Grimson, K. Farh, W. Johnston, P. Garrett- Engele, L. Lim, D. Bartel. Micro RNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell.* 27(1): 91-105, 2007.

[17] J. Hakenberg, S. Bickel, C. Plake, U. Brefeld, H. Zahn, L. Faulstich, U. Leser, T. Scheffer. Systematic feature evaluation for gene name recognition. *BMC Bioinformatics*, 6(1):S9, 2005.

[18] S. Kim, J. Nam, J. Rhee, W. Lee, and B. Zhang. MiTarget: miRNA target gene prediction using an SVM. *BMC Bioinformatics*, 7:411, 2006.

[19] R. Lee, R. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75: 843-854, 1993.

[20] B. Lewis, C. Burge, and D. Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are Micro RNA Targets. *Cell*, 120:15-20, 2005.

[21] C. Lu, K. Kulkarni, F. Souret, R. Valliappan, S. Tej, R. Poethig, I. Henderson, S. Jacobsen, W. Wang, P. Green, B. Meyers. MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res.* 16(10):1276-88, 2006.

[22] C. Lu, S. Tej, S. Luo, C. Haudenschild, B. Meyers. Green Elucidation of the small RNA component of the transcriptome. *Science* 309: 1567-1569, 2005.

[23] K. Lynn, A. Fernandez, M. Aida, J. Sedbrook, M. Tasaka, P. Masson, M. Barton. The PINHEAD/ZWILLE gene acts pleiotropically in Arabidopsis development and has overlapping functions with the ARGONAUTE1 gene. *Development.* 126:469, 1999.

[24] N. Markham and M. Zuker. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Research.*, 33: W577-W581, 2005.

[25] Meyers Laboratory. Arabidopsis MPSS plus. *http://mpss.udel.edu/at/*, 2007.

[26] B. Meyers, T. Vu, S. Tej, M. Matvienko, H. Ghazal, V. Agrawal, C. Haudenschild. Analysis of the transcriptional complexity of Arabidopsis by massively parallel signature sequencing. *Nature Biotech*. 22: 1006-1011, 2004.

[27] E. Murchison and G. Hannon. MiRNAs on the move: MiRNA biogenesis and the RNAi machinery. *Curr. Opin. Cell Biol*, 16: 223–229.

[28] C. Napoli, C. Lemieux, R. Jorgensen. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. Plant *Cell*. 2: 279-289, 1990.

[29] A. Oppenheim and R. Schafer. *Discrete-Time Signal Processing*, 1999.

[30] A. Pasquinelli. MicroRNAs: deviants no longer. *Trends in Genetics,* 18(4):171-173, 2002.

[31] R. Rajagopalan, H. Vaucheret, J. Trejo, D. Bartel. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana. Genes & Dev*. 20: 3407-3425, 2006.

[32] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich. Fast and effective prediction of micro RNA/target duplexes. *RNA,* 10:1507-1517, 2004.

[33] B. Reinhart, F. Slack, M. Basson, A. Pasquinelli, J. Bettinger, A. Rougvie, H. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature,* 403: 901–906, 2000.

[34] B. Reinhart, E. Weinstein, M. Rhoades, B. Bartel, and D. Bartel. MicroRNAs in plants. *Genes & Development,* 16:1616-1626, 2002.

[35] M. Rhoades, B. Reinhart, L. Lim, C. Burge, B. Bartel, and D. Bartel. Prediction of plant miRNA Targets. *Cell*, 110(4):513-520, 2002.

[36] K. Robinson-Beers, R. Pruitt, C. Gasser. Ovule Development in Wild-Type Arabidopsis and Two Female-Sterile Mutants. *Plant Cell.* 4: 1237, 1992.

[37] C. Rock, Q. Luo, V. Stolc, M. Samanta. Evidence for upstream antisense transcription associated with targets of miRNAs in Arabidopsis. Plant Biology. Manuscript in Revision.

[38] D. Schwarz and P. Zamore. Why do miRNAs live in the miRNP? *Genes & Development,* 16:1025-1031, 2002.

[39] V. Stolc, M. Samanta, W. Tongprasit,H. Sethi, S. Liang, D. Nelson, A. Hegeman, C. Nelson, D. Rancour, S. Bednarek, E. Ulrich, Q. Zhao, R. Wrobel, C. Newman, B. Fox, G. Phillips, J. Markley, and M. Sussman. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. In Proceedings of the National Academy of Sciences, 102(12):4453-4458, 2005.

[40] A. Van der Krol, L. Mur, M. Beld, J. Mol, A. Stuitje. Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. Plant *Cell*. 2:291-9, 1990.

[41] X. Wang, J. Reyes, N. Chua, and T. Gaasterland. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biology*, 5(9):R65. 2004.

[42] K. Yamada, J. Lim, J. Dale, H. Chen, P. Shinn, C. Palm, A. Southwick, H. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. Quach, M. Tripp, C. Chang, J. Lee, M. Toriumi, M. Chan, C. Tang, C. Onodera, J. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, Akiko Enju, Andrew D. Goldsmith, Mani Gurjal, Nancy F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. Hsuan, K. Iida, M. Karnes, S. Khan, E. Koesema, J. Ishida, P. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. Davis, A. Theologis, and J. Ecker. Empirical Analysis of Transcriptional Activity in the *Arabidopsis* Genome. *Science*, 302(56460):842 – 846, 2003.

[43] Y. Zhang. MiRU: an automated plant miRNA target prediction server. *Nucleic Acids Research*, 34: W451–W454, 2005.

PERMISSION TO COPY

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Texas Tech University or Texas Tech University Health Sciences Center, I agree that the Library and my major department shall make it freely available for research purposes. Permission to copy this thesis for scholarly purposes may be granted by the Director of the Library or my major professor. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my further written permission and that any user may be liable for copyright infringement.

Agree (Permission is granted.)

Viktoria Gontcharova 11-28-2007

_____ _____

Student Signature Date

Disagree (Permission is not granted.)

_____ _____

Student Signature Date