SYSTEMATIC COMPUTATIONAL ANALYSIS OF POTENTIAL RNA INTERFERENCE REGULATION IN *Toxoplasma gondii*

A Thesis Submitted to
The Graduate School of Engineering and Sciences of
Izmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

In Molecular Biology and Genetics

by
Mehmet Volkan ÇAKIR

December 2009
İZMİR

We approve the thesis of **Mehmet Volkan ÇAKIR**

Assist. Prof.Dr. Jens ALLMER
Supervisor

Assoc. Prof.Dr. Bilge KARAÇALI
Committee Member

Assist. Prof.Dr. Bünyamin AKGÜL
Committee Member

15 December 2009

Assoc. Prof. Dr. Sami DOĞANLAR
Head of the Department of Molecular Biology and Genetics

Assoc. Prof. Dr. Talat YALÇIN
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGMENTS

I am heartily thankful to my supervisor, Assist.Prof.Dr.Jens Allmer, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of my subject and of scientific research process in general. This study would not have been possible without his mentoring and support.

I also would like to make a special reference to Assoc.Prof.Dr.Bilge Karaçalı and Assist.Prof.Dr.Bünyamin Akgül for their invaluable suggestions and discussion.

I should express my gratitude to TÜBİTAK for financial support during my M.Sc. education.

I cannot express how much I am grateful to my wife Gözde Selin Çakır and to my family who have always been there for me no matter what the situation is.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

ABSTRACT

SYSTEMATIC COMPUTATIONAL ANALYSIS OF POTENTIAL RNA INTERFERENCE REGULATION IN *Toxoplasma gondii*

RNA-mediated silencing was first described in plants and became famous by studies in *Caenorhabditis elegans*. RNA interference (RNAi) is the mechanism through which an RNA interferes with the production of other RNAs in a sequence specific manner. MiRNAs are a type of RNA which originate from the genome with their active form being ss-RNAs of 21-23 nucleotides in length. They are being transcribed as pri-miRNAs then processed in the nucleus by Drosha to pre-miRNAs with a stem-loop structure and ~70 nucleotides in length. This stem-loop containing pre-miRNAs is then processed in the cytoplasm to ds-RNA one strand of which will serve as interfering RNA.

Toxoplasma gondii is a species of parasitic protozoa which causes several diseases. *T.gondii* emerges as a good candidate for computational efforts with its small genome size, publicly available genome files and extensive information about its gene structure, either based on experimental data or the prediction with several gene finders in parallel. Therefore, it seems important to establish the regulatory network composed of RNAi which may be beneficial for the Toxoplasma community.

Within this context the pool of possible stem-loop constitutive transcripts are produced, further analysis of this pool for desired 2D structure is integrated and mapping of possible RNAi regulation to *T.gondii*'s genome is established. In connection with computational assessment and mapping, the derived information is provided as a database for quick lookup using a convenient web interface for experimental studies of RNAi regulation in *Toxoplasma*, thus reduce time and money costs in such studies.

ÖZET

Toxoplasma gondii'DE POTANSİYEL RNA İNTERFERANS REGÜLASYONUNUN SİSTEMATİK SAYISAL ANALİZİ

RNA vasıtasıyla gen anlatımının kontrolü ilk olarak bitkiler üzerinde yapılan çalışmalarda tanımlanmış ve *Caenorhabditis elegans* üzerindeki çalışmalar ile yaygınlaşmıştır. RNAi, bazı özel RNA'ların belirli oranda tamamlayıcılık gösterdiği diğer RNA'ların üretimini etkilemesi ve üretim basamaklarına müdahelesi mekanizmasına verilen isimdir. Mekanizmanın bu dizisel tamamlayıcılık özelliği dizi analizi yardımıyla işlemsel analizi mümkün kılmaktadır.MiRNA'lar en son aktif formlarında yaklaşık 21-23 nükleotid uzunluğunda olan ve canlıların kendi genomlarından kaynaklanan tek zincirli RNA'lardır. MiRNA'ların transkripsiyonları pri-miRNA denilen RNA dizileri şeklinde yapılır ve hücre çekirdeğinde nükleaz Drosha enzimi tarafından bir kesim işlemi ile yaklaşık ~70 nükleotid uzunluğunda pre-miRNA adı verilen sap-ilmik(stem-loop) yapılarına dönüştürülürler. Daha sonra bu sap-ilmik yapısındaki pre-miRNA'lar sitoplazmada bir diğer enzim kesim işlemi ile bir zinciri aktif miRNA olarak aktivite gösterecek olan çift zincirli RNA'lara dönüştürülür.

Toxoplasma gondii pek çok hastalığa neden olan bir protozoan parazitidir. Küçük boyutlu genomu ve bilim insanlarının kullanımına sunulan genom dosyaları ile *T.gondii* işlemsel çalışmalar için iyi bir aday olarak görünmektedir. *T.gondii* yüksek ökaryotlara benzer bazı özellikler sergilemesine rağmen *T.gondii*'de RNAi regülasyonu ile ilgili kapsamlı bilgi bulunmamaktadır. Bu nedenle *T.gondii* için oluşturulacak RNAi kontrol ağı *T.gondii* konusunda çalışan bilim insanları için çok önemlidir.

Bu bilgilerin ışığında çalışmamızda *Toxoplasma* genomundan muhtemel sap-ilmik yapıları oluşturulmuş, arzulanan sap-ilmik yapısına sahip olan diziler seçilmiş ve bu dizilerden kaynaklanan aktif miRNA'lar Toxoplasma genomuna haritalanmıştır. Bu sonuçlara bağlı olarak elde edilen bilgi pratik bir web arayüzü ile deneysel *Toxoplasma* çalışmaları yütütecek araştırmacıların hizmetine sunulması ve *Toxoplasma* çalışmalarının zaman ve para olarak maliyetini düşürmesi amaçlanmaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

ABBREVIATIONS

T. gondii	*Toxoplasma gondii*
RNAi	RNA interference
dsRNA	Double stranded RNA
PTGS	Post transcriptional gene silencing
ssRNA	Single stranded RNA
mRNA	Messenger RNA
siRNA	Small interfering RNA
rasiRNA	Repeat-associated short interfering RNAs
piRNA	PIWI interacting RNA
dsRBD	dsRNA binding domain
miRNA	micro RNA
RdRP	RNA-dependant RNA polymerization
Pri-miRNA	Primary micro RNA
TU	Transcription unit
RNase	Ribonuclease
DGCR8	DiGeorge syndrome chromosomal region 8
Dcl	Dicer-like
Pre-miRNA	Precursor mi-RNA
tasiRNA	Trans-acting siRNA
DUF	Domain of unknown function
PIWI	P-element induced wimpy testis
PPD	Paz Piwi domain
PACT	protein kinase R (PKR)-activating protein
TRBP	TAR RNA-binding protein
TAR	transactivating response
miRISC	miRNA-containing RISC
VM	Virtual machine
Shreps	Shape representatives
E.histolytica	*Entamoeba histolytica*
G.intestinalis	*Giardia intestinalis*

CHAPTER 1

INTRODUCTION

1.1. *Toxoplasma gondii*

T.gondii, which belongs to the Apicomplexan phylum, is a parasite to warm-blooded animals. It is estimated to infect about one-third of the world's human population and can cause plenty of syndromes such as encephalitis, chorioretinitis and congenital infection and myocarditis (Kim and Weiss 2008, Blader and Saeij 2009). *T.gondii*, was first discovered in 1908 by Nicolle and Manceaux in tissues of *Ctenodactylus gundi* a hamster-like rodent (Nicolle and Manceaux 1908). For following years *T. gondii*-like organisms were found in various species. Finally, viable *T. gondii* was isolated and proved to be identical with its human isolate by (Sabin and Olitsky 1937). Its complex life cycle comprises of two phases: a sexual cycle in its feline definitive hosts and an asexual cycle in its intermediate hosts (Figure 1.1) (Dubey 2004). The transmission between hosts occurs via three reported mechanisms: Congenital, through carnivorism, fecal-oral (Dubey 2008).

Apicomplexan parasites belong to a phylum which consists of diverse and highly specialised organisms (Meissner et al. 2007). Recent efforts in the field of apicomplexans, completion of certain genomes, analyses of transcriptome and proteome of these parasites have provided invaluable and comprehensive insights into the the parasite (Carlton et al. 2002, Gardner et al. 2002, Kooij et al. 2006, Meissner et al. 2007). By these significant progresses the objective now become to identify new candidates for the development of new drugs and new possible therapies. For all these efforts to be fulfilled a wise bioinformatics and statistical analysis has to be made first.

Studies on *T.gondii* can be considered as momentous for three reasons. Primarily *Toxoplasma* can cause severe diseases especially in developing fetuses and in immune-compromised patients besides current available drugs, that cannot act against chronic *Toxoplasma* infections, are poorly tolerated and come with severe side effects while there are cases that resistance to these drugs are reported (Blader and Saeij 2009, Aspinall et al. 2002, Baatz et al. 2006, Dannemann et al. 1992). Secondly, while

availability of reverse genetic tools exhibits its significance on apicomplexan biology publicly available genomic data makes *Toxoplasma* an attractive organism for all fields of molecular biology (Meissner et al. 2007). The availability of cost-effective DNA sequencing methods has revolutionized *T. gondii* research as well as other sequenced organisms by making genomic information available (Kim and Weiss 2008). Experimental and genomic data related to *T.gondii* are put together and published at the single organism database ToxoDB (Kissinger et al. 2003, ToxoDB 2009). Thirdly, *Toxoplasma* is used as a model system for other parasitic Apicomplexans (Ajioka 1997, Blader and Saeij 2009).

Pathogenesis caused by apicomplexan parasites is reported to be as a result of an unbounded parasite biomass expansion accompanied by tissue destruction and inflammation (Gubbels et al. 2008). Spreading of the infection through tissues to the organ systems, which can lead to death in the situations of weak immune response, is mainly provoked by the renewal of parasitic invasion, replication and cell lysis cycles (Gubbels et al. 2008). High growth rates seem to be crucial to virulence and are a major cause of severe infections. Observed correlation between the magnitude of progeny, the rate of multiplication and the intensity of disease in malaria supports this knowledge on mechanism of infection (Chotivanich et al. 2000, Dondorp et al. 2005, Reilly et al. 2007, Timms et al. 2001). Responsibility of growth rate in highly pahogenic infections is proved to be pivotal in mice in recent studies (Radke et al. 2001, Taylor et al. 2006). This virulence and growth rate relation emphasizes the clinical and therapeutical importance of understanding RNAi regulation in apicomplexians through the study of McRobert and McConkey 2002 .

Cell division mechanism of *T. gondii* is characterized by the relationship between mitosis and cytokinesis which is assured by the unique apicomplexan internal budding mechanism (Striepen et al. 2007).

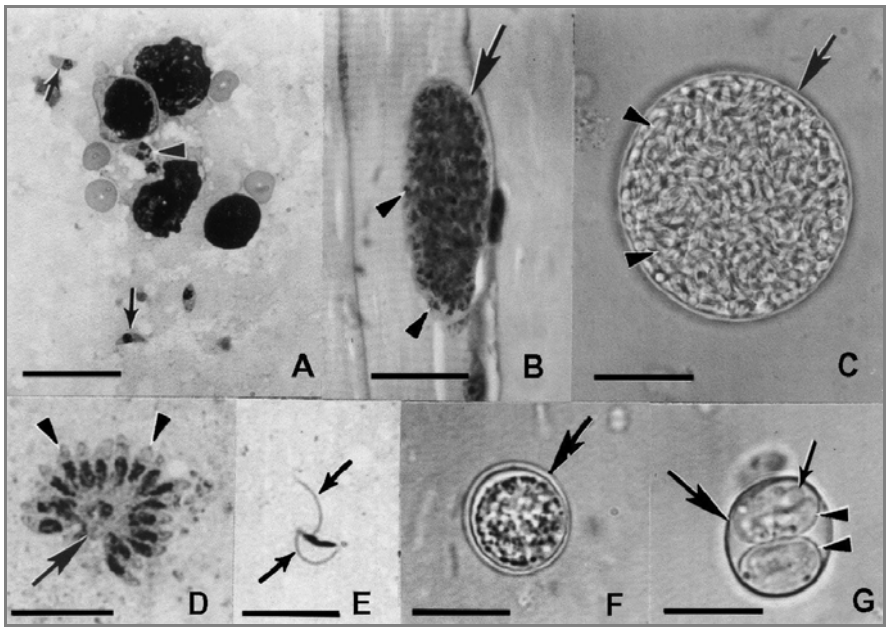

Figure 1.1 Stages of *Toxoplasma gondii*. (A) Tachyzoites in impression smear of lung. Note crescent-shaped individual tachyzoites (arrows), dividing tachyzoites (arrowheads) compared with size of host red blood cells and leukocytes (B) Tissue cysts in section of muscle. The tissue cyst wall is very thin (arrow) and encloses many tiny bradyzoites (arrowheads) (C) Tissue cyst wall (arrow) and hundreds of bradyzoites (arrowheads) (D) Schizont (arrow) with several merozoites (arrowheads) (E) A male gamete with two flagella (arrows) (F) Unsporulated oocyst in faecal float of cat feces. Double-layered oocyst wall (arrow) enclosing a central undivided mass. (G) Sporulated oocyst with a thin oocyst wall (large arrow), two sporocysts (arrowheads). Each sporocyst has four sporozoites (small arrow) which are not in complete focus (Source: Hill and Dubey 2002).

High throughput analysis tecniques regarding Toxoplasma biology have become feasible with the availability of genome sequence and genomic information of Toxoplasma. Studies of gene expression is one of the major studies on Toxoplasma. The gene expression patterns of development is studied using gene expression analysis and DNA microarrays (Radke et al. 2005, Boyle et al. 2006). The improved technology, rise of new techniques and tools in molecular biology, genomics, epigenetics, epigenomics, proteomics and metabolomics provided more advanced and comprehensive studies. These studies are reviewed in study by K. Kim and Weiss (Kim and Weiss 2008).

1.2. RNAi

1.2.1. Definition and Scope

The term RNA interference (RNAi) refers to a cellular process by which a double-stranded RNA (dsRNA) down-regulates the expression of a gene/genes in a sequence specific manner. It has also known as posttranscriptional gene silencing (PTGS) and is a field of enormous current interest. RNA interference develops early in the eukaryotic lineage and plays essential roles in many diverse regulatory mechanisms such as cellular immunity, modulation of chromatin structure, and development (Baulcombe 2004, Mello and Conte 2004). The discovery of RNA interference has been widely regarded as a major breakthrough in modern molecular biology with its regulatory roles and dramatic effects.

RNA interference was first discovered in plants and approved to occur in a wide variety of eukaryotic organisms (Meister and Tuschl 2004). The silencing of genes in a sequence specific manner via the injection of dsRNA in Caenorhabditis elegans lead to coining of term RNA interference (Fire et al. 1998). RNA silencing mechanisms were first thought to be antiviral mechanisms which serves as a protection mechanism against RNA viruses or regulates the random integration of transposable elements (Meister and Tuschl 2004). But significant and general role of RNAi in the regulation of gene expression became clear by realization of specific regions that encode RNAs which can fold on themselves to produce double stranded hairpins(Ambros 2004).

RNAi mechanism is initiated by dsRNA precursors that vary in length, origin and structure while expressing some certain perceptible similarities. DsRNA precursors end up with short RNA duplexes after a couple of processing steps ~22 nucleotides in length which then guide recognition, cleavage or translational repression of complementary single-stranded RNAs (ssRNA), such as messenger RNAs (mRNA) or viral genomic/antigenomic RNAs (Meister and Tuschl 2004). It is also reported that RNAi mechanism interferes with chromatin modification as well (Lippman and Martienssen 2004). Biogenesis and functions of RNAi have been reviewed extensively (Bushati and Cohen 2007, Filipowicz et al. 2008, Rana 2007). The tissue specific regulation of miRNAs in development emphasizes the significance of RNAi regulation

(Houbaviy et al. 2003, Lagos-Quintana et al. 2001, Lagos-Quintana et al. 2002, Lau et al. 2001, Mourelatos et al. 2002, Pasquinelli et al. 2000).

Three types of naturally occurring small RNAs have been described due to their functions: short interfering RNAs (siRNAs), repeat-associated short interfering RNAs (rasiRNAs) and microRNAs (miRNAs) (Meister and Tuschl 2004). However there are other specific types: heterochromatic siRNAs, transacting siRNAs (ta-siRNAs), natural antisense siRNAs (nat-siRNAs), and, in metazoans, the Piwi-interacting RNAs (piRNAs) (Meins et al. 2005, O'Donnell and Boeke 2007, Vaucheret 2006, Vazquez et al. 2004, Zhang et al. 2006). PiRNAs are shown to prevent the spreading of selfish genetic elements by methyl-dependent epigenetic silencing and cleavage of transposon mRNA (Aravin et al. 2007, Brennecke et al. 2008, Klattenhoff and Theurkauf 2008, Nowotny and Yang 2009). Nevertheless they can be reduced to two major classes according to their origin: siRNAs and miRNAs (Chapman and Carrington 2007, Zamore and Haley 2005).

Both RNAi mechanisms require endonucleolytic cleavage of dsRNA to generate approximately 20–30 base pairs (bp) dsRNA with two nucleotide overhangs at 3’ ends of both strands (Nowotny and Yang 2009). It is a stepwise process catalysed by dsRNA-specific RNase-III-type endonucleases, known as Drosha and Dicer. These RNases contain two major domains: a catalytic RNase III and a dsRNA-binding domains (dsRBDs) (Meister and Tuschl 2004).

MiRNA pathway starts in the nucleus with the processing of pri-miRNAs by Drosha to produce pre-miRNAs, hairpin structures approximately 70 nucleotides in length, which are then transported to the cytoplasm to be processed by another RNAse Dicer into miRNA duplexes while siRNAs are generated by Dicer from dsRNA precursors (Nowotny and Yang 2009). The miRNA precursors are stem-loop structure forming noncoding transcripts with characteristic bulges and mismatches within the folded molecule which are thought to destabilize miRNA precursors and sets forth important features for processing (Hutvagner and Zamore 2002, Khvorova et al. 2003). After Dicer process, one strand of short dsRNA duplexes is incorporated into the RISC complex (RNA-induced silencing complex) ,which is multiprotein complex with ability to incorporate ssRNA and slicer function, for targeting mRNAs by base pairing (Nowotny and Yang 2009).

DsRNAs can be produced either by RNA-dependant RNA polymerization (RdRP) or by association of transcripts which exhibits a certain amount of similarity.

Production of dsRNA by RdRP can be initiated by viruses while transposon derived dsRNAs can be an example of production via repetitive sequences (Meister and Tuschl 2004). Such mechanisms end up with constitution of regulatory assciations siRNAs or rasiRNAs, that often trigger the degradation of mRNA and/or modification of chromatin (Meister and Tuschl 2004). Precursors of other class of RNAi regulation agents, miRNAs, is produced by folding of transcripts, which contain a region of complementary inverted repeat 20 to 50 base pair in length, to form stem-loop and hairpin structures (Meister and Tuschl 2004). MiRNAs can mediate translational repression and/or mRNA degradation. These known properties of RNAi elements brought one of the most effective tools of modern molecular biology: introduction of long dsRNAs or siRNAs into the cells to inactivate gene expression (Meister and Tuschl 2004).

Most miRNAs are conserved in closely related species while they have homologs in distant species. Approximately a third of *C. elegans* miRNAs seem to have homologs in humans (Lim et al. 2003) suggesting that their functions could also be conserved throughout the evolution which encourages new efforts on diverse organisms in this area. Also the low sequence conservation in the loop compared to miRNA segment is reported (Kim and Nam 2006).

In animal miRNA mechanism partial complementarity between miRNA and 3' untranslated region (UTR) of target mRNA, like those of *Caenorhabditis elegans*, often leads to translational inhibition while in plants, miRNAs mostly lead the cleavage of sequence-complementary mRNAs (Meister and Tuschl 2004, Lu et al. 2008).

RNAi biogenesis is defined as a mixture of recognition and cleavage. Biogenesis of RNAi has became clear by structural studies of proteins and their complexes involved, such as Argonaute, PIWI, RNase III, Dicer, Drosha, and DGCR8. There are repeated functions in RNAi mechanism; namely, recognition of the 3'-end and 5'-end of RNA, binding of dsRNA, and cleavage of dsRNA at a defined distance from one end (Nowotny and Yang 2009).

1.2.2. Metabolism

1.2.2.1. Sources and Expression

As detailed above RNAi can be divided into two major classes according to their sources: endogenous and exogenous RNAis. Exogenous RNAi can be elicited by dsRNA supplement from outside the cell, endogenous RNAi can be elicited from transcription of coding or noncoding genomic sequences (Allen et al. 2005, Ambros 2004, Baulcombe 2004, Grishok et al. 2005, Lippman and Martienssen 2004, Mello and Conte 2004, Peragine et al. 2004). Endogenous RNAis are the microRNAs, which function in the regulation of gene expression in multicellular eukaryotes, exogenous RNAis are short RNA duplexes which function in a variety of transcriptional and post-transcriptional gene-silencing processes (Lee et al. 2006).

MiRNA genes can be categorized according to their locations in transcription units (TUs): intronic miRNAs in protein coding TUs, intronic miRNAs in noncoding TUs and exonic miRNAs in noncoding TUs (Kim and Nam 2006). Analyses of miRNAs put forward that location of the majority of mammalian miRNA genes(~70%) reside in defined transcription units and most of these miRNA(~70%), that dwell in transcription units, are found in introns which indicates previous mining efforts that exclude TUs might have missed some miRNA locations (Rodriguez et al. 2004, Kim and Nam 2006).

Small interfering RNAs(siRNAs) are generated from long dsRNA such as viruses, transgenes, transposons or artificially introduced by members of a family of endoribonucleases, called Dicers, which contain an aminoterminal RNA helicase domain, a central ‘PAZ’ motif, and carboxy-terminal tandem ribonuclease III domains (Zamore 2004). MiRNAs are processed from transcripts called pri-miRNAs, which contain self-complementary inverted sequences, to approximately 70 nucleotide long pre-miRNAs by the activity of Rnase Drosha and then thse pre-miRNAs are transported out of the nucleus by Exportin-5 (Bohnsack et al. 2004, Lund et al. 2004, Yi et al. 2003). MiRNA biogenesis is initiated via transcription by RNA polymerase II (Cai et al. 2004, Kim 2005, Lee et al. 2003, Lee et al. 2002). Also there is an alternative way through production of miRNAs. This alternative pathway for miRNA biogenesis is driven by debranched introns which mimic the structural features of pre-miRNAs and

called ‘mirtrons (Ruby et al. 2007). Despite their distinct sources and different processes they possess similarities in their functions (Lu et al. 2008).

1.2.2.2. Processing

1.2.2.2.1. Drosha Cleavage

MiRNAs originate from peculiar stem-loop and hairpin structures in primary transcripts (pri-miRNAs), which contain both 5’ cap and poly(A) tail, through two censecutive RNase III-mediated cleavages. Drosha cleaves next to the lower stem matches on the hairpin structure and Dicer cleaves near the loop to generate a miRNA:miRNA* duplex (Lee et al. 2003, Tomari and Zamore 2005). Rnase-III proteins are dsRNA-specific endonucleases which are grouped into three classes according to their domain organization and cuts both strands of dsRNAs in a staggered manner while leaving 2 nucleotide overhangs on 3’ end both strands as a characteristic of this family of enzymes (Lee et al. 2003). Bacteria and yeasts possess class I Rnases that contain a conserved RNase-III domain and a dsRNA-binding domain (dsRBD) (Figure 1.2). Drosha falls into class II of RNases that contain a tandem RNase-III domain and one dsRBD and an extended amino-terminal domain with and unknown function (Figure 1.2) (Filippov et al. 2000, Fortin et al. 2002). Dicer and its homologues reside in class III with a helicase/ATPase domain,a DUF283 domain a PAZ domain, tandem nuclease domains and a dsRBD (Figure 1.2) (Han et al. 2006, Lee et al. 2003). Nevertheless different classifications are proposed one of which divides RNase III enzymes into two classes: class I contain enzymes which contain a single RNase III domain and function as homodimers, and class II contains enzymes which have two catalytic RNase III domains and act as monomers (Jaskiewicz and Filipowicz 2008). Essentially RNase III enzymes’ catalytic domains are well conserved while they reasonably exhibit similar mechanism of action (Han et al. 2006).

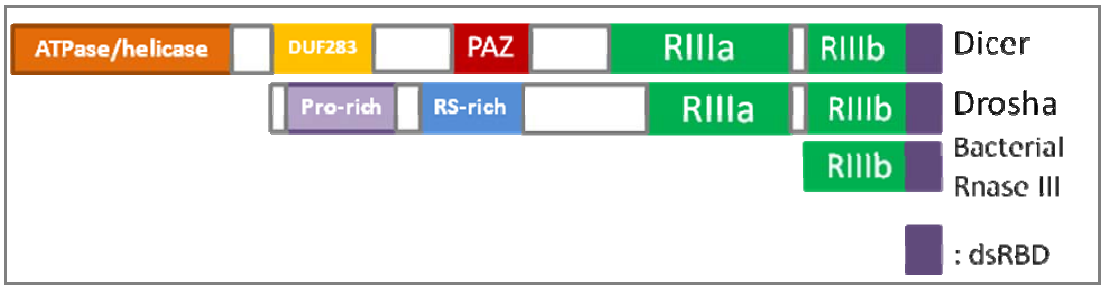

Figure 1.2 Domains of three classes of the RNase III family proteins represented by human Dicer (class 3), human Drosha (class 2), and bacterial RNase III (class 1) (Source: Zhang et al. 2004).

In the nucleus pri-miRNAs (Figure 1.3) which are generated by special miRNA genes or derived from introns are trimmed by the RNase III Drosha to ~70 nucleotide long premiRNAs. These pre-miRNAs are then transported to cytoplasm by a dsRNA-binding protein Exportin-5 and are cleaved by Dicer to act as the functional miRNAs (Bohnsack et al. 2004, Ketting et al. 2001, Kurreck 2009, Lee et al. 2003, Lee et al. 2002, Lee et al. 2004, Lund et al. 2004, Yi et al. 2003). However the length of Drosha products vary in different species; ~80 nt in animals and more variable in plants (Kim and Nam 2006, Ullu et al. 2004).

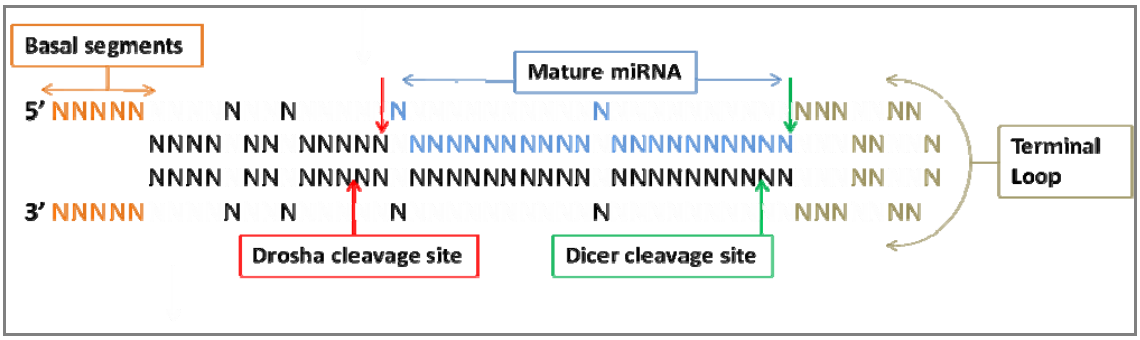

Figure 1.3 An average pri-miRNA can be divided into four parts: a terminal loop, the upper stem, the lower stem, and basal segments. Red and Green arrows indicate cleavage cites of Drosha and Dicer (Source: Han et al. 2006).

Drosha cleavage is accrued in nucleus in a protein complex called microprocessor which contains a dsRNA-binding protein, DGCR8 (DiGeorge syndrome chromosomal region 8, known as Pasha in *Caenorhabditis elegans* and *Drosophila melanogaster* (Denli et al. 2004, Gregory et al. 2004, Han et al. 2004, Landthaler et al. 2004). Yeast two hybrid screening and immunopurification tests indicated the

interaction between Drosha and DGCR8/Pasha, furthermore neither of them have been observed to be active during pri-miRNA cleavage while their association reproduce the active complex (Denli et al. 2004, Gregory et al. 2004, Han et al. 2004, Landthaler et al. 2004). This coordinated association of factors set forth the necessity to DGCR8 for the Drosha cleavage to be accomplished (Gregory et al. 2004, Han et al. 2004).

Pri-miRNA processing consists of two sequential steps, namely substrate recognition and catalytic reaction. Initially, DGCR8 anchors at the SD(ssRNA-dsRNA) junction, which is the boundary between dsRNA region and ssRNA region by interacting with the stem and generates a precleavage complex for intervention of Drosha (Han et al. 2006). Drosha does not directly hook up with RNA before generation of this precleavage complex. Right after this precleavage phase, the dsRBD of Drosha interacts with the stem and adjust the location of processing center at ~11 bp from the SD junction (Han et al. 2006). There is no defined sequence motif for Drosha which indicates that Drosha recognizes some shared structural motifs in hairpins (Han et al. 2006).

Pri-miRNA formation and processing of them into pre-miRNAs are crucial steps in miRNA biogenesis since they designate the sequence that Dicer acts on and essentially pre-expose embedded miRNA sequences via defining one border of the sequence (Han et al. 2006).

A typical animal pri-miRNA contains a stem of ~33 bp, a terminal loop of ~10 bases and various number of flanking nucleotides (Figure 1.3). Although the general approximate semblance of pri-miRNAs can be constructed into four parts via thermodynamical stability analysis with approximate lengths assigned to these parts; mismatching basal segments (varying in length), well aligned lower stem (~11 nt), well aligned upper stem (~22 nt) and terminal loop (varying in length) (Zeng et al. 2005). Basal segments vary in length, however it is shown that Drosha can process efficiently the pri-miRNAs that contain only ~20 nt outside the cleavage sites (Han et al. 2004).

The location of the Drosha processing center, cleavage site, is fundamentally determined by the distance from the SD junction while flanking sequences seem to be crucial for processing and the loop shows the slightest effect during processing (Lee et

al. 2003, Yekta et al. 2004, Zeng and Cullen 2005). It is shown by directed mutation analysis that the cleavage site of Drosha is ~11 bp from the SD junction (Han et al. 2006). Length l deletions in lower stem shifts cleavage cite l nucleotides (Han et al. 2006). Neither the alteration of the region between cleavage site of Drosha and terminal loop via replacement, deletion or insertion nor the length of the terminal loop changes the cleavage site of Drosha supporting that basal segments are crucial while terminal loop is not vital for Drosha process (Han et al. 2006).

1.2.2.2.2. Dicer Cleavage

The characteristics of RNaseIII enzymes and the similarities between their products made researchers, one of whom is Brenda Bass, aware of the RNaseIII traces on elements of RNAi (Bass 2000, Moss 2001). Not long after such an enzyme, Dicer, is identified to have a role in RNAi pathway in Drosophila (Bernstein et al. 2001).

Dicer is a member of the RNase III family of dsRNA specific nucleases which cleaves RNAs with their signature, 3' flanking 2 nucleotide overhangs, left behind (Bernstein et al. 2001, Ji 2008).

Many organisms have Dicer homologs including *C. elegans*, *Arabidopsis thaliana*, *Schizosaccharomyces pombe*, worms, flies, fungi and humans. Dicer is found in the cytoplasm of nearly all eukaryotic cells and acts by recognizing the 5' and 3'ends of dsRNA and cleaving a specific distance from that end to 21– to 28 nt siRNAs or microRNAs (Elbashir et al. 2001, Hutvagner et al. 2001). Dicer functions in association with other proteins and ions, like Mg^{2+}, in multiprotein complexes and *in vivo* (Forstemann et al. 2005, Jaskiewicz and Filipowicz 2008, Saito et al. 2005). The length of Dicer products are told to be enough to provide adequate sequence complexity to uniquely identify a single gene in an eukaryotic genome (Macrae et al. 2006). Thus, they are quite versatile tools for translational regulation. Dicer carries out an auxiliary role other than cleavage activity by loading RNA products into multiprotein RNA-induced silencing complexes (RISC) (Gregory et al. 2005, Maniataki and Mourelatos 2005, Pham et al. 2004, Tomari et al. 2004). Loaded RISC uses its burden as a guide to act and interfere gene expression through certain mechanisms; mRNA degradation (Hammond et al. 2001), translational inhibition (Pillai et al. 2005), and heterochromatin formation (Verdel et al. 2004).

Diverse species can encode for different number of Dicer/Dicer-like enzymes; humans and *C.elegans* encode only one, *Drosophila* encodes two, and the Arabidopsis encode for four different Dicer enzymes (Lee et al. 2004). The four different plant Dcls(Dicer like enzymes) have different roles. Dcl-1 enzyme processes pri-miRNAs and the precursor miRNA (pre-miRNAs); Dcl-2 copes with siRNAs which are related to antiviral defense mechanism, Dcl-3 generates siRNAs that are involved in chromatin modification and transcriptional silencing while Dcl-4 generates trans –acting siRNAs (tasiRNAs) that originate from non-coding RNAs and interfere with the expression of their target mRNAs (Borsani et al. 2005, Giraldez et al. 2006, Kurihara and Watanabe 2004, Park et al. 2002, Vazquez et al. 2004, Xie et al. 2005, Xie et al. 2004). The distinct functions of Drosophila Dicers are pre-miRNA processing for Dicer-1 and siRNA production for Dicer-2 (Jaskiewicz and Filipowicz 2008).

All RNAi pathways require Dicer while in some cases presence of Drosha is not essential and Drosha cleavage is not prerequisite of RNAi biogenesis (Bernstein et al. 2001, Ruby et al. 2007).

There are several common features and domains generally shared between higher eukaryotes. Five prevalent domains in metazoan and plant Dicer enzymes are an ATPase/helicase domain, a DUF283 domain, a PAZ domain, two RNase III domains and a dsRBD (Jaskiewicz and Filipowicz 2008, Moss 2001). In addition lower eukaryotes usually expose a less complex domain organization. Structural information about function of Dicer derived from the crystal structure of a fully active Dicer from *Giardia intestinalis* (Macrae et al. 2006). This enzyme of *G.intestinalis* naturally contains only the PAZ and two catalytic domains (Macrae et al. 2006). These two domains associate to form an 'internal dimer' and linked to PAZ domain from the opposite end of the molecule (Blaszczyk et al. 2001, MacRae et al. 2007). The 65-A° distance between the PAZ domain and the catalytic domain active sites introduces the *G.intestinalis* Dicer product length that span around 25–27 base pairs suggesting that catalytic domain and PAZ domain determines the length of product in a ruler like fashion (Figure 1.4). Furthermore, deletion studies proved this measuring mechanism of Dicer (MacRae et al. 2007). PAZ domain measures dsRNAs from 3' end which enables production of longer RNAs than expected from substrates containing a 5' extension (MacRae et al. 2007).

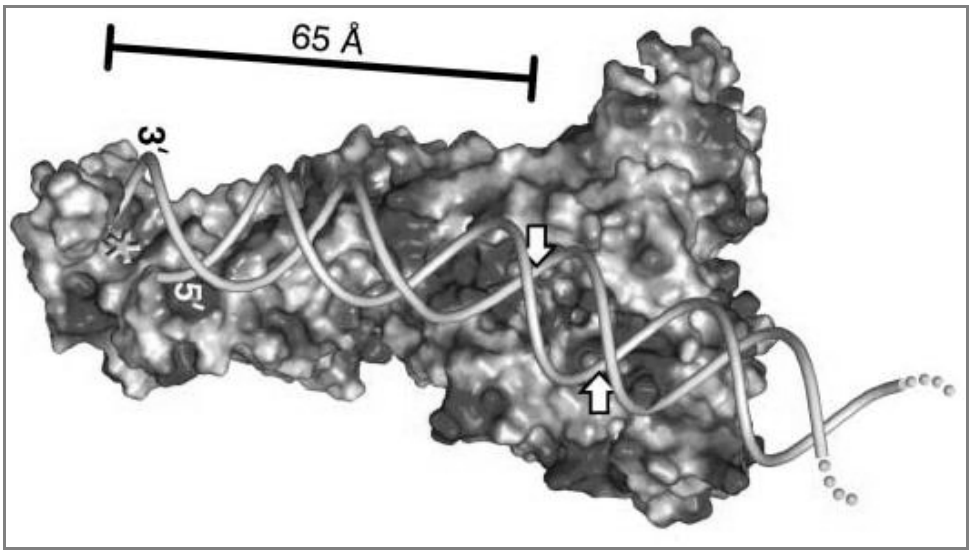

Figure 1.4 Front and side views of a surface representation of Giardia Dicer with modeled dsRNA. The distance of PAZ domain from processing center displays the length of Dicer products which is 25-27 nucleotides in *G.intestinalis* (Source: Macrae et al. 2006).

It is shown that an open helical end is essential for thorough processing (MacRae et al. 2007). However terminal blocking of dsRNAs result in an internal cleavage with reduced kinetics which restores the normal kinetics via uncovering 2-nt 3′-overhang-containing (Zhang et al. 2002). This polar interaction and processing mechanism of Dicer is carried out by RNaseIIIa and RNaseIIIb domains (Zhang et al. 2004). RNaseIIIa interacts and processes 3' end while RNaseIIIb interacts and processes 5' end. The exact product size is determined either by PAZ domain and the structure of substrate (MacRae et al. 2007, Rose et al. 2005). Besides, different features of substrates, like recessed 3' end, can end up with slightly different lengths of products (MacRae et al. 2007). By the guidance of these structural features, a model which proposes the binding of Dicer to 3' ends of dsRNA via PAZ domain and positioning is introduced (Figure 1.5) (Zhang et al. 2004).Recent evidences about RNase III proteins are in agreement with this model (Gan et al. 2006).

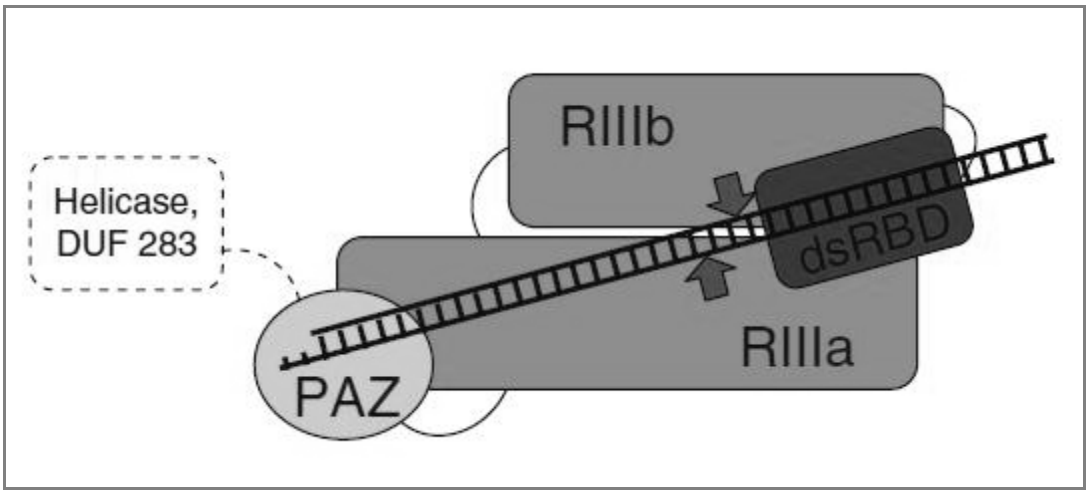

Figure 1.5 A model of dsRNA processing by human Dicer (Zhang et al. 2004). Individual domains of Dicer are shown in different colors. The enzyme contains a single dsRNA cleavage center with two independent catalytic sites. The center is formed by intramolecular dimerization of the RNase IIIa and RNase IIIb domains. The placement of the RIIIa domain illustrates the fact that this domain cleaves the 3′-OH-bearing and protruding RNA strand. DsRBD positioning is arbitrary (Source: Jaskiewicz and Filipowicz 2008).

The complex functional units of Dicer expose diverse roles for its domains. The PAZ, dsRBD, and RNase III domains of Dicers are related to dsRNA binding and cleavage (Jaskiewicz and Filipowicz 2008). The PAZ and PIWI domains act together in PPDs (Paz Piwi domain proteins) and involve in RISCs. PPD family proteins are characterized by a central PAZ domain and a carboxy terminal PIWI domain (Carmell et al. 2002). The interaction of Dicer with PPDs is shown and in this context PPDs are thought to be required for loading of siRNAs/miRNAs to RISCs (Hammond et al. 2001, Sasaki et al. 2003). Furthermore it is reported that PAZ domains are not essential for PPD proteins to interact with Dicer (Tahbaz et al. 2004).

Up until this phase of miRNA biogenesis there are functions that are repeatedly required: recognition of the ends of RNA, binding of dsRNA, and cleavage at a defined distance from one end (Nowotny and Yang 2009). Nature uses some functional modules to handle these tasks. PAZ domain recognizes the 3'-end of RNA in both miRNA and siRNA, the Mid domain of PIWI and Ago binds the 5'-phosphate, dsRNA-binding domains which are found in Drosha, DGCR8, and Dicer binds dsRNA, an Rnase III-like endonuclease domain (endoND) performs all dsRNA cleavage in RNAi (Nowotny and Yang 2009). Slicer activity is carried out by the Rnase H resembling PIWI domain in PIWI and Ago which can nick mRNA (Nowotny and Yang 2009).

Dicer cleavage produces double stranded miRNA:miRNA* duplexes one strand of which will be integrated into a protein complex called RISC.

1.2.2.2.3. RISC Assembly and Strand Selection

RISC (RNA induced silencing complex) is a multiprotein complex which binds one strand of RNAi elements (siRNA or miRNA) to catalyze a sequence specific association with its target mRNA. SiRNA and miRNA duplexes have to be unwound and separated into single strands prior to their assembly into RISC where they cause the sequence specificity of the silencing activity (Preall and Sontheimer 2005). Production of double stranded RNA duplexes is the common feature between pathways of miRNA and siRNA. Antisense strands that are incorporated into RISC are called small interfering (si) RNAs if they are perfectly complementary with their target and if they arise from long dsRNA. They are called microRNAs if their annealing is imperfect with their target and if they arise from pre-miRNA hairpins with a loose definition and crude classification (Preall and Sontheimer 2005). However, there is no limitation for miRNAs to display perfect complementarity with their targets or that siRNAs can not act if they do not expose perfect complementarity with their targets. Besides, the specifity of an interaction diminishes with decreasing number of matches between RNAi element and its target.Besides reduced length of partial matches between an RNAi active element and its target also reduces its specifity.

One strand of the RNA duplex product generated by Dicer is called guide strand while the other is the passenger strand (Berkhout and Jeang 2007). The guide strand is incorporated into RISC while the passenger strand is degraded. The resemblance between siRNA:siRNA* and siRNA:mRNA duplexes provokes the passenger strand to be the primary target of active RISC in the siRNA RISC loading step (Matranga et al. 2005, Rand et al. 2005).

The RISC, which contains a member of the Argonaute protein family as a core component, has been purified from fly and human cells (Hammond et al. 2001, Hutvagner and Zamore 2002, Martinez et al. 2002). Argonaute family proteins are approximately 100 kDa proteins referred to as PPD proteins which emphasizes two shared domains; the PAZ and the PIWI domain (Cerutti et al. 2000). Moreover, Dicer-2 (DCR-2) is the enzyme which is required for cleaving long precursor dsRNAs into

siRNAs in *Drosophila* (Bernstein et al. 2001, Lee et al. 2004). DCR-2 and R2D2, a protein with a dsRNA-binding domain; interact during Dicer processing (Liu et al. 2003). DCR-2-R2D2 association is also required for loading of siRNAs onto the (RISC) (Liu et al. 2003) hence it is named the RISC-loading complex (RLC) (Lee et al. 2004, Tomari et al. 2004). The siRNA duplex, which is processed by DCR-2-R2D2, is passed onto the endonuclease Argonaute2 (AGO2) that cleaves target mRNAs (Tolia and Joshua-Tor 2007). AGO2 is also responsible for the cleaveage of the passenger strand (Leuschner et al. 2006, Matranga et al. 2005, Miyoshi et al. 2005, Rand et al. 2005). Furthermore, the cleavage of the passenger strand is a necessary and crucial step governing RISC activity in the siRNA pathway (Matranga et al. 2005, Rand et al. 2005). On the other hand miRNAs can also be incorporated into RISC by a by-pass mechanism which is called cleavage independent pathway (Matranga et al. 2005). Essentially, it is proposed that the base pairing quality at the seed region (between nucleotide positions 2 and 8) is a key determinant for the selection between cleavage dependent or cleavage independent pathway (Matranga et al. 2005). Still, the exact mechanism of miRNA duplex unwinding and the factor(s) that are responsible for unwinding remain unknown. However, it is known that the orientation of guide strand in the RISC is assured by Dicer interacting proteins: PACT with its dsRNA binding domain and TAR RNA-binding protein (TRBP) (Chendrimada et al. 2005, Gatignol et al. 1991, Lee et al. 2006).

Additionally, recent studies in *D. melanogaster* established that sequence asymmetry and strand instability have an effect on the selective strand processing of pre-miRNA hairpin precursor and siRNA duplexes (Schwarz et al. 2003). Together, recent studies on unwinding propose a correlation between the thermodynamic stability profiles of siRNAs and miRNAs and their RNA interference stimulating ability.

Upon Dicer processing and before unwinding, two strands of miRNAs, which are aforementioned passenger and guide strands, still exist in their duplex form and they should be unwound prior to miRISC (RISC containing miRNA) formation. Strand rejection in miRNA mechanism is not accompanied by miRNA* strand cleavage by AGO1 (Matranga et al. 2005). Still it has been shown that *Drosophila* AGO1 proteins play roles in miRNA processing and sequential translational repression (Miyoshi et al. 2005). Through these processes, miRISC is formed and RISC is activated.

An miRISC mediates miRNA function(s) inside cells. In plants, miRISC often requires perfect complementarity between miRNAs and their targets which are mostly

3’ untranslated regions of mRNAs for proper functioning (Berkhout and Jeang 2007). In this perfect complementarity manner, plant miRISC mediates mRNA cleavage and consequent degradation similar to the function of si-RISC - mediated silencing (Elbashir et al. 2001, Fire et al. 1998). On the other hand animal miRISC and mRNA interaction somehow can tolerate a certain level of mismatches at the particular positions and the function is mostly dependent on matches in the 5’ seed region (the region between the 2nd and 7th nucleotide following the 5’ end ofmiRNAs (Lewis et al. 2005). Furthermore, these mismatches in animal miRNAs are also thought to be effective in blocking endonucleolytic cleavage of animal mRNAs.

Once an miRISC-mRNA interaction forms this interaction either represses translation or mediates premature mRNA decay (Berkhout and Jeang 2007). Current data suggests several mechanisms regarding miRISC-mRNA interaction: inhibition of translational initiation, increasing co-translational degradation of nascent proteins, reducing the elongation rate of translation, increasing the rate of mRNA deadenylation (Humphreys et al. 2005, Maroney et al. 2006, Nottrott et al. 2006, Petersen et al. 2006, Pillai et al. 2005, Wu et al. 2006). Recent data shows a distinct mechanism in which eIF6 component of miRNA-RISC prevents productive assembly of 80 S ribosome complex (Chendrimada et al. 2007). However, the question which conditions result in which mechanism remains in debate. Nonetheless, P-bodies, which are ribosom-free translationally silent cytoplasmic organelles, are suggested to be involved in silencing mechanisms via arresting miRISC-mRNA associations (Liu et al. 2005).

It has been shown that during miRISC formation reduced internal stability of the 5’terminus and overall low stability across the initially unwound region are conserved properties of miRNA hairpin precursors (Khvorova et al. 2003). SiRNAs are derived from several sources such as viral infections, transgene activity, or transposons (Zamore 2004). Moreover, processing of siRNA and miRNA precursors involves common cellular proteins. In convenience with the similarity between miRNA and siRNA biogenesis it has been shown that functional siRNAs exhibit low internal stability at the 5’ terminus, as miRNAs, suggesting similar unwounding principles (Khvorova et al. 2003). Furthermore, major stability differences are observed within siRNA duplexes at other positions than at the terminal end only. Functional siRNAs possess molecuels with low internal stability while nonfunctional siRNAs are enriched with high internal stability (Khvorova et al. 2003). Moreover, it has been suggested that flexibility in the 9-14 region might serve to proper target cleavage and product release upon RISC

mediated cleavage by RISC associated endonuclease and RISC* regeneration (Khvorova et al. 2003). Internal stability profiles of functional siRNAs/miRNAs might play important roles in several steps of the RNAi regulation such as duplex unwinding, strand selection, and product release. For unwinding and RISC loading to occur efficiently duplexes need to be destabilized either by external agents or by low internal stability of the sequence itself (Khvorova et al. 2003). These destabilizing elements are base pair mismatches, gaps, and bulges in miRNAs. Convenient with their prevalent perfect siRNA:target complementarity siRNAs can not contain such destabilizing elements. Thus, they have low stability which originates from sequence itself at key positions. Indeed, it has been shown by in vitro studies that different absolute and relative stabilities of the base pairs at the 5' ends of siRNA strands determine which strand will participate in the RNAi pathway (Schwarz et al. 2003). It has also been shown in flies that asymmetry of the siRNA duplex causes preferential binding of the Dcr-2/R2D2 protein to more stable end of siRNA, by this way introducing asymmetric strand incorporation into the Ago2-RISC complex (Tomari et al. 2004). However, loading of RISC by different, Dcr-2/R2D2 independent mechanisms, is possible for miRNAs. There are miRNAs which are loaded to Ago1-RISC by a mechanism that is independent of Dcr-2/R2D2 (Tomari et al. 2007). Also, some recent studies on miRNA expression profiles have shown that the relative expression levels of the two strands may vary widely among tissues suggesting that loading mechanism and conditions that affects this mechanism have not been explicitly identified yet (Ro et al. 2007). Remarkably, in some tissues thermodynamically unfavourable miRNA strands are observed to cause interference at levels comparable to or greater than their thermodynamically more favourable siblings.

Upon RISC* formation, multiple turnovers is needed for repetitive functioning and effective silencing of RISC (Khvorova et al. 2003). Every turnover includes loading, target cleavage at the position opposite the center of the guide antisense strand, dissociation of target molecule, product release and reassociation. Observed stability profiles of functional RNAi elements at special positions may facilitate these processes and by this way allow RISC* to asociate with subsequent substrate mRNA strands (Khvorova et al. 2003).

1.2.2.3. Effects and Regulation Mechanisms of miRNAs

MiRNAs carry out their functions as components of ribonucleoprotein complexes (RNPs) or RNA-induced silencing complexes which are called micro-ribonucleoproteins (miRNPs) and miRNA-induced silencing complexes (miRISCs).

In plants miRNAs often cause mRNA cleavage when a perfect complementarity between miRNA and its target is given (Filipowicz et al. 2008, Jones-Rhoades et al. 2006). In metazoans miRNAs usually repress translation via mostly imperfect base pairing between miRNA and target (Brennecke et al. 2005, Doench and Sharp 2004, Grimson et al. 2007, Lewis et al. 2005, Nielsen et al. 2007).

Eukaryotic translation can be considered in three phases which are initiation, elongation and termination. Initiation step starts with the recognition of the mRNA terminal cap structure (Kapp and Lorsch 2004, Merrick 2004). Recruitment of eukaryotic initiation factors (eIFs: eIF4E, eIF4F, eIF4G, eIF3) and poly-adenylate binding protein 1 (PABP1) facilitates the recruitment of 40S ribosomal subunit and constitutes a circular mRNA structure via bringing two ends of mRNA close together which is thought to be effective in ribosome recycling (Derry et al. 2006, Kapp and Lorsch 2004, Merrick 2004, Wells et al. 1998). However all translation initiation mechanisms do not rely on cap recognition. In those cases recruitment of 40S ribosomes and translation initiation is carried out by interaction of internal ribosome entry sites (IRESs) (Jackson 2005). Elongation of the translation starts with the association of 60S ribosome subunit. MiRNAs can be effective in different stages of translation.

It has been shown that miRNAs repress the translation of properly capped mRNAs and do not repress the translation of irregular capped or IRES containing mRNAs suggesting that miRNAs interfere with the function of eIF4E (Humphreys et al. 2005, Pillai et al. 2005). Another model connects miRNA interference with translation initiation. This model suggests direct effect of miRNAs via competing of AGO proteins with eIF4E for cap binding (Kiriakidou et al. 2007). Also miRNA binding site containing mRNAs are deadenylated even if they contain a proper cap and despite their not containing IRES (Wang et al. 2006). Thus it is known that miRNAs disrupt the relation between cap and poly(A) tail and abolish mRNA circularization.

MicroRNAs do not only act on the initiation phase of translation. A recent study suggests a role on 60S subunit association. The factor eIF6 accompanies 60S subunit

from nucleolus to cytoplasm and is needed for proper 60S subunit biogenesis (Basu et al. 2001, Sanvito et al. 1999, Si and Maitra 1999). Partial deletion of eIF6 helps mRNAs to escape from miRNA regulation which confirms interference of miRNAs in 60S subunit association (Chendrimada et al. 2007). Other studies suggest premature termination of translation while repression in post-initiation and elongation steps are presented by cosedimentation of miRNAs with polysomes (Kim et al. 2004, Nelson et al. 2004, Petersen et al. 2006, Vasudevan and Steitz 2007).

The exact repression mechanisms of the initiation or the post-initiation step is not thoroughly known, however, there are proposed mechanisms, like slowing down or even stalling the elongation, depending on previous translational post-initiation repression studies (Mootz et al. 2004, Ruegsegger et al. 2001).

Recent studies suggest that miRNA mediated repression is accompanied by mRNA deadenylation, destabilization and and mRNA decay (Bagga et al. 2005, Behm-Ansmant et al. 2006, Giraldez et al. 2006, Wu and Belasco 2005, Wu et al. 2006). Eukaryotic mRNA degradation can be either by 5'→3' degradation by exonuclease XRN1 or 3'→5' degradation by exosomes (Parker and Song 2004). Both pathways are initiated by poly(A) tail shortening and are controlled by recruitment of decaying components. Recruitment of the decay machinery results in mRNPs and leads to deadenylation and decapping. Degradation of mRNAs mostly takes place in special structures known as P-bodies which are enriched in translational repressors, mRNA deadenylation, decapping, and degradation enzymes (Eulalio et al. 2007, Parker and Sheth 2007). The *D. melanogaster* P-body protein GW182 and its homologues in mammalians and worms seem to interact with AGO and PIWI domains and protect miRNAs from decay (Behm-Ansmant et al. 2006, Ding et al. 2005, Meister et al. 2005, Till et al. 2007). Moreover, translational repression can be result of mRNA destabilization accordingly tissue specific repression/destabilization and different degrees of destabilization have been observed (Mishima et al. 2006, Schmitter et al. 2006).

It is anticipated that miRNAs control translation through several mechanisms yet it is too early to draw a comprehensive and complete picture of miRNA regulation in cells. Nonetheless, there are proposed models of AGO mediated cap-dependent translational inhibition followed by either mRNA degradation, proteolyis of nascent polypeptides and mRNA destabilization. Still the question under which conditions

which repression pathway takes place needs further efforts to be clarified. Also it is crucial to identify all factors involved,all RISC components, and signalling pathways.

1.2.3. RNAi Regulation in *T.gondii*

Unfortunately, there is a limited number of experimental studies on RNAi regulation in apicomplexan parasites. However, studies indicate presence of RNAi regulation in apicomplexans. In spite of the debates on RNAi metabolism in *T.gondii* there are studies which suggest an RNAi regulation in *T.gondii* that resembles the one of eukaryotes (McRobert and McConkey 2002). Moreover, it has been reported that the genome of *Toxoplasma gondii* contains candidate sequences with convincing similarity to RNAi genes (Ullu et al. 2004). Database mining studies predicts that *Entamoeba histolytica* and *Giardia intestinalis* have an RNAi pathway . The *G.intestinalis* case supports appearance of dsRNA mediated gene silencing early in the evolution of eukaryotic lineage (Ullu et al. 2004). Also existence of an inducible RNAi system in lower eukaryotes has been proposed by several studies (Bastin et al. 2000, Shi et al. 2000, Wang et al. 2000).

A study on *Plasmodium falciparum* suggests the presence of RNAi regulation in apixomplexan parasites. However, in this study it is the effect of RNAi but not the RNAi elements which cause that effect shown (Malhotra et al. 2002). Another study on *Plasmodium berghei* identified not only the effect but also the effective elements as siRNAs which are verified to be effective short RNAs with supplementary efforts (Mohmmed et al. 2003). However genome mining studies failed to identify any homologues of Dicer, Piwi, Paz and RdRp in Plasmodium databases (Ullu et al. 2004).

Studies on RNAi regulation in *T. gondii* do not show much divergence from efforts on other apicomplexans. One encouraging study has shown downregulation of uracil phosphoribosyltransferase (UPRT) via introduction of dsRNAs (Al-Anouti and Ananvoranich 2002). Notwithstanding, the lack of identified effective siRNAs or miRNAs in *T. gondii*, mining of the *T. gondii* genome revealed existence of putative ORFs (open reading frames) which resemble several genes present in RNAi pathways (Ullu et al. 2004). These RNAi genes are potential homologues of AGO, Dicer and RdRp. The potential *T. gondii* AGO-like protein contains both a PAZ and a Piwi

domain (Ullu et al. 2004). Additionally, it has been pointed out that potential RNAi genes of *T. gondii* might function in the production of miRNAs (Ullu et al. 2004).

The potential RNAi property of two protozoas, *G.intestinalis* and *E.histolytica*, have been shown by either database mining or isolation of the active Dicer enzyme, respectively (Macrae et al. 2006, Ullu et al. 2004). However, a comprehensive and fully functional RNAi regulation remains elusive and in debate (Meissner et al. 2007). Thus, computational efforts on RNAi regulation may provide benefitial information to the scientific community of these organisms.

1.2.4. Aim of the Study

The essence of our study is to acquire a collection of potential miRNA interferences from *T. gondii*. Their sources and their potential targets as well as all potential miRNA and target sequences involved. Type of origin (whether they are exon, intron or non-coding sequences) and their locations in the genome of *T.gondii* is also of prime interest. By this way, the type and location of sequences which can potentially act as miRNAs and their potential can be inferred. This mapping will be presented in a database along with all for this study relevant calculated values that belong to source sequences, target sequences and their interaction. The values are minimum free energy values of hairpins, length of the arms of hairpins, number of mismatches in hairpins, length of the longest stretch in hairpins, mismatches in the longest stretch, length of lower stem matches in hairpins, length of the terminal loop, minimum free energy values of miRNA duplexes, number of different mononucleotides, dinucleotides and trinucleotides in miRNAs, number of matches, mismatches and gaps in the miRNA:target interactions, e-values of the interaction. The derived information will enable the Toxoplasma community and scientists of RNAi field to have a potential RNAi regulation map of *T.gondii* at hand. Moreover, provided values can be used to assess the validity of interactions and can aid in determining thresholds for future RNAi interaction maps.

On the other hand the system developed in this study can be adjusted to be used on other genomes since it provides users with enough flexibility via its switches and comands. Users can adjust almost all of the thresholds and fine tune the system according to their taste. Furthermore, implemented classes can be used seperately in

other systems of miRNA regulation analysis or the overall system can be used by just enriching its implementation with custom classes, methods or filtering steps since a genome-wide RNAi analysis entails a great amount of time and coding effort.

A genome-wide analysis involves many thresholds, values, decision and/or filtering steps. The constructed system for genome-wide RNAi analysis can be used to design a file format for RNAi analysis since there is still no specific format for RNAi analysis.

CHAPTER 2

MATERIALS AND METHODS

2.1. Materials

2.1.1. Programming Language

In order to run a genome wide analysis it is essential to have an automated system. Using a pre-implemented system or a pre-designed program is not a judicious way of fulfilling the step by step RNAi regulation analysis on ~80 mb length *T.gondii* genome. Besides, there is no standart genome wide analysis program or system for *T.gondii*. In spite of programs we include in our system in certain steps we wrote our own code. Implementations are written in Java™ programming language from Sun Microsystems.

Java™ is an high level object-oriented programming language with comprehensive internal libraries and an object oriented nature. It is originally developed by James Gosling at Sun Microsystems and released in 1995 as a core component of Sun Microsystems' Java™ platform. One of the most significant advantageous of Java™ language is portability. Portability means, programs written in the Java™ language run similarly on any supported hardware/operating-system. Java™ language code is compiled to Java bytecode, which is an intermediate representation, instead of directly to platform-specific machine code. This feature of Java brings portability with the need for a virtual machine(VM) which is written specifically for the host hardware and interpret Java™ bytecode instructions. Another useful property of Java™ is reusability. Modularly written and sorted basic methods and functional modules can be integrated in other systems and can easily be reused in other projects.

2.1.2. Java™ : Classes, objects, fields, methods and inheritance

Owing to the object oriented nature of Java™ it is fundamental to have a basic understanding of object oriented concepts to implement any design in java and/or grasp

the very idea of any system in Java™. Our system consists of classes each of which undertakes a crucial step in our artificial miRNA biogenesis. Thus a certain amount of familiarity is essential to understand the structure and design of our system.

The idea of object oriented programming emerges from real world objects and basic concepts have been established in 70's and 80's (Bobrow and Winograd 1977, Goldberg and Robson 1983, Minsky 1974, Weinreb 1981). There were more than fifty object oriented programming languages with very limited spread in 1986 (Stefik and Bobrow 1986). Nevertheless, object oriented programming concept became prevalent and widely used with the introduction of C++ and later Java ™. There are some fundamental notions that lie at the heart of object oriented logic, which are objects, classes, fields, methods and inheritance.

Objects in the object oriented logic are entities that abstractly resemble to real world objects. Books, cell phones, automobiles, birds, televisions, pencils etc… Abstract interpretation of real world objects entails notions and rules. Within this respect real world objects share two fundamental characteristics: *states* and *behaviours* (Sun Microsystems 2009). States basically refer to the properties of objects while behaviours refer to acts of or on objects. "States" and "behaviours" of a real world object can be derived by just observing its physical properties and acts. A modest object as a TV can have several states which correspond to its *fields* in object oriented programming: colour of its cover, its open/closed state, the current channel if it is open, level of its volume, level of its contrast, level of its colour, number of channels in its memory etc… And acts of a TV object which correspond to its *methods* in object oriented programming: open, close, switch channel, open specific channels, adjust volume, adjust colour, adjust contrast etc. The number of relevant fields and methods can be increased arbitrarily. However, a good object oriented design should be wisely constructed to prevent inconsistencies. A software object can be visualized by considering its field as its core and its methods as its crust which establishes connection with other objects (Figure 2.1) (Sun Microsystems 2009).

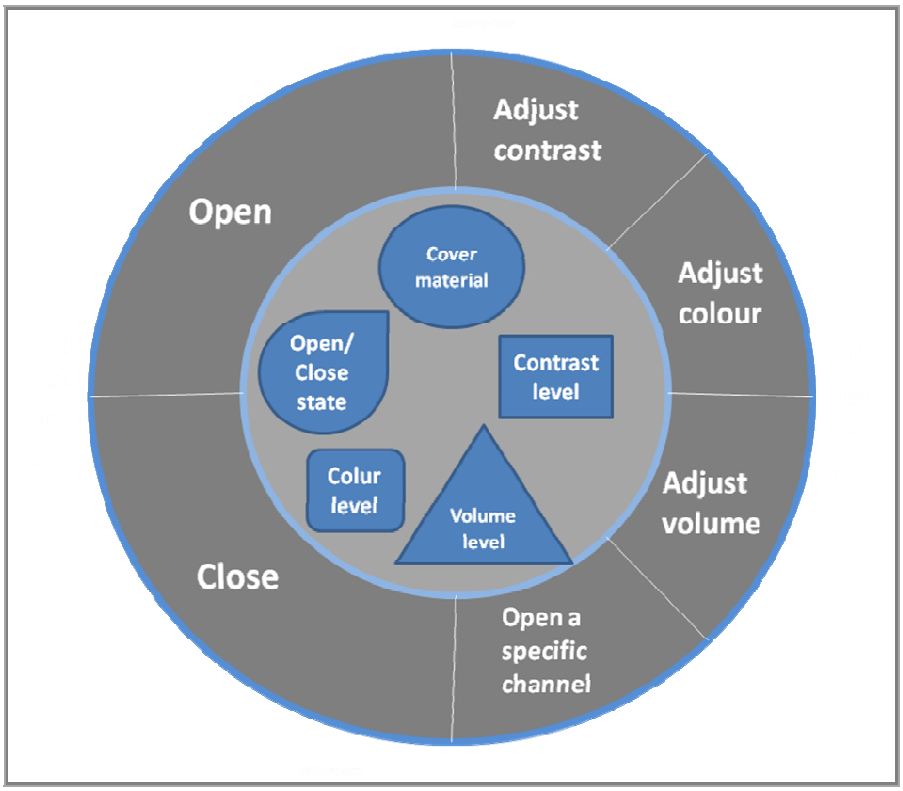

Figure 2.1 Visualization of a software object (representing a TV). The core of the object represents its fields (states of the abstracted real world object) and the cover of the object represents its methods (behaviours of the abstracted real world object) (Source: Sun Microsystems 2009).

Object oriented design and programming provides several fundamental advantages which are among others:

- modularity by providing ability to write and maintain source code of an object separate from the source code of other objects
- information hiding by keeping details of the implementation hidden and seperated from utilization of objects
- code reuse by providing the ability to use your or others objects in several projects
- debugging ease due to ease of capturing problematic objects (Sun Microsystems 2009).

Real world objects are not unique entities. There might be millions of TVs and perhaps thousands of TVs with same model and setup as your TV. It is not fallacious to

say they share attributes and behaviours to a certain level. In object oriented concept your TV is an *instance* of the *class of objects* called TVs (Sun Microsystems 2009). Classes are the *blueprints* from which relevant objects can be created (Sun Microsystems 2009).

Inheritance emphasizes common properties of a class of objects. Plasma TVs, LCD TVs, flat TVs all share common characteristics of televisions even if they have certain level of difference (Figure 2.2). Likewise introns, exons, non-coding regions have properties in common: they all consist of four lettered alphabet, they all have length, location etc. Object oriented concept enables objects to inherit common properties and behaviours from ther classes while the class which presents inherited properties and behaviours become *superclass* and the class which inherits properties and behaviours become *subclasses* (Sun Microsystems 2009).

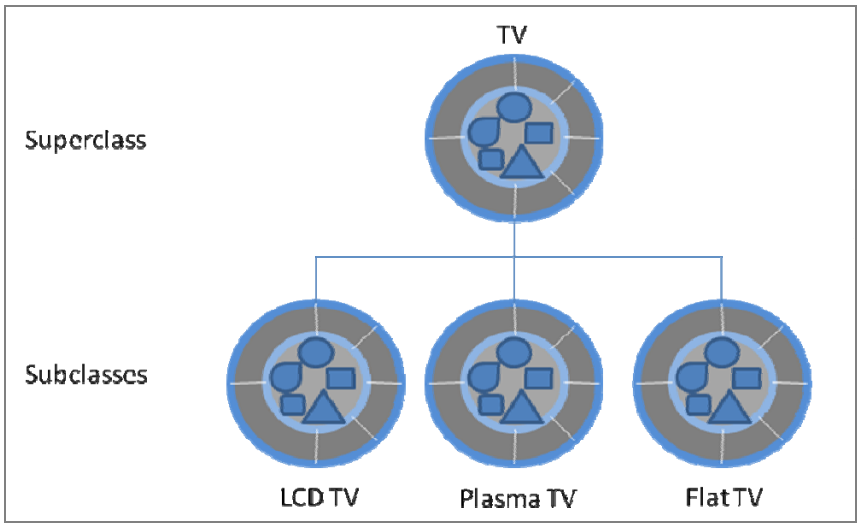

Figure 2.2 Superclasses, subclasses and inheritance
(Source:(Sun Microsystems 2009).

2.1.3. File types

Our analysis starts with two types of files regarding *T. gondii* genome. One of these files is in FASTA format, the specifications of which are published in NCBI database, and the other is in GFF format which is defined and specified by Sanger Institute. Both files are obtained from ToxoDB – Toxoplasma gondii Genome Resource

(Kissinger et al. 2003). ToxoDB is a single organism database which containes publicly available -both annotated and raw- files of *T. gondii* genome.

2.1.3.1. Genomic Sequence File of *T.gondii* in Fasta Format

FASTA format is an easily parsable, widely used by bioinformatics applications and simple structured format of plain text files. A sequence in FASTA format begins with an identifier line which is followed by lines of sequence data. The identifier line is indicated by a greater-than sign (“>”) at its beginning and distinguished from sequence data via this character (Figure 2.3). Identifier lines – not necessarily- contains some information like database name, organism name, length of the sequence, a primary key which uniquely designates that particular sequnce… etc. Structure of identifier lines is something that is to be described by the producer of FASTA file. It can differ due to purposes of files.

Usually all of the lines in FASTA files are shorter than 80 characters in length however there is no such limitation about lengths of lines. Still there is a consensus that identifier lines are shorter than 80 characters and sequence lines are shorter than 60 characters in length.

Both nucleic acid sequences and protein sequences can be stored in fasta format as long as they are represented in the standard IUB/IUPAC amino acid and nucleic acid codes (IUPAC-IUB commission on biochemical nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents 1971).

Genomic FASTA file of *T. gondii* - TgondiiME49Genomic_ToxoDB-5.0.fasta- which we used in our study belongs to the ME49 strain and obtained from 5.0 release of ToxoDB. It contains 328 identifier lines and their corresponding sequences and 65.132.538 bytes in size.

```
>gb|DS984777|organism=Toxoplasma_gondii_ME49|version=2008-07-23|
length=62811
TGCCCCCCTCACCATGTCACGACACTCCGGCTCCGCTCCCTGTTTCACCTTCTGTGCCCT
ACCGCAGCTGCCAACACGGGCTCATCGGATGCAATGAAACCAACAACGCCTGACAGAGTA
CTGGCGACTGCAAACGAGGGCTACCATGACGCACCCATCATCAACCATTGCAGGGCC...
..........................................................

>gi|16443|emb|CAA78106.1| protein kinase [Arabidopsis thaliana]
MDKYDVVKDLGAGNFGVARLLRHKDTKELVAMKYIERGRKIDENVAREIINHRSFKHPNI
IRFKEVILTPTHLAIVMEYASGGELFDRICTAGRFSEAEARYFFQQLICGVDYCHSLQIC
HRDLKLENTLLDGAPAPLLKICDFGYSKSSILHSRPKSTVGTPAYIAPEVLSR.......
..........................................................
```

Figure 2.3 Examples of DNA and protein sequences in fasta format from *T.gondii* and *A.thaliana* respectively.

2.1.3.2. Genomic Feature File of *T.gondii* in Gff Format

GFF (General Feature Format) is an easy to parse structured format of plain text files like FASTA. GFF Protocol Specification was initially proposed by Richard Durbin and David Haussler with amendments proposed by Lincoln Stein, Suzanna Lewis, Anders Krogh and others. Its latest default stable version is 2.0. Although GFF is not intended to be used for complete data management of the analysis and annotation of genomic sequence, its simple structure and low complexity representation of information is extremely valuable to conserve consistency and robustness of genomic analysis process.

Information is stored line by line in GFF files, every line contains information regarding different sequence. Every line contains eight tab departed columns and additional two fields –attributes and comments- from version 2.0 onwards (Figure 2.4). These features of format are defined by Sanger Institute but like FASTA format they are not inviolable rules. They are subject to manipulations with respect to intentions. Aforementioned eight columns in GFF files are as follows; seqname, source, feature, start, end, score, strand and frame. The field seqname stands for the name of sequence, name of the chromosome or some other sequence related name optionally. Source field stands for the name of the database name or the source (public database, an annotation program etc…) from which the sequence is originated. Feature field holds the information about the interested feature of the sequence like exon, intron, gene, mRNA and so forth. New features can be defined and used freely, there is no prohibition or

restriction about use of these fields so GFF is a relaxed and mutable format. Start and end fields stands for start and end indexes of sequences. Sequence numbers start from 1, so the start index must be an integer greater than 1 and smaller than end index. Also end index must be bigger than start index. Reasonably end index minus start index gives the length of the sequence. Score holds a floating value for any value needed: it can be a calculated probability value which designates presence of an interested pattern, minimum free energy value calculated for folded structure of a sequence or anything else needed. If there is no relevant score '.' character is used instead. Strand field indicates the location of sequence. It can be '+' for plus strand, '-' for minus strand or '.' if there is no relevant strand information. Frame column stands for the position of open reading frame. There are three possible positions for reading frame if only one strand is considered since codons consist of three nucleotides, with two strand there are six possible reading frames. Frame field can hold '0', '1', '2' or '.'. '0' indicates that the first base of sequence corresponds to the first base of a codon. '1' indicates that the second base of the sequence corresponds to the first base of a codon, and '2' means that the third base of the sequence is the first base of a codon. If there is no relevant strand information frame field is set to '.'. Attribute field is an optional field to use if needed. It contains semicolom departed any relevant information. '#' character is reserved for comments field to maintain parsability and comments can be used anywhere in the overall file.

Genomic GFF file of *T. gondii* - TgondiiME49_ToxoDB-5.0.gff- which we used in our study belongs to the ME49 strain and obtained from 5.0 release genomic sequence files of ToxoDB regarding *T. gondii*ME49 strain. It is 86.763.142 in bytes. It contains chromosome names in the field of seqname, source name –which is ApiDB (Aurrecoechea et al. 2007) - in the source field, mRNA, CDS, gene or exon in the feature field, start and end indexes of sequences in respective fields, no score in score field, strand information in strand field, frame information in frame field and some related information in attribute field which also contains unique identifiers of sequences. Unique identifiers are keys that consist of a character string and uniquely indicates one and only one sequence.

```
TGME49_chrIb ApiDB exon 1010888 1013419 . - . ID=apidb|TGME49_009050-1..
TGME49_chrIb ApiDB exon 1010278 1010497 . - . ID=apidb|TGME49_009050-2..
TGME49_chrIb ApiDB exon 1009672 1009874 . - . ID=apidb|TGME49_009050-3..
TGME49_chrIb ApiDB gene 1018944 1026367 . + . ID=apidb|TGME49_009060....
TGME49_chrIb ApiDB mRNA 1018944 1026367 . + . ID=apidb|TGME49_009060-1..
.....................................................................
.....................................................................
```

Figure 2.4 The structure of *T.gondii* TGME49 gff file release-5.3 (Source: ToxoDB 2009).

2.1.4. Programs Included in the System

Although we design and implement our own method, we used some external and auxiliary scientific programs to achive some straightforward steps in our system. While they reduce the flexibility of the system they provide proper execution and swiftness. Pre-implemented programs in our system for three individual steps are RNAshapes for folding RNA sequences into stem-loop and hairpin structures (Giegerich et al. 2004, Reeder and Giegerich 2005, Steffen et al. 2006, Voss et al. 2006), RNAhybrid for calculating free energy of two associated RNA sequences (Kruger and Rehmsmeier 2006) and BLAST for exhausting whole genome with our antisense strands (Altschul et al. 1990).

2.1.4.1. RNAshapes

RNAshapes is freely available software package which integrates three RNA analysis tools, namely the analysis of shape representatives, the calculation of shape probabilities and the consensus shapes approach (Steffen et al. 2006). Source code and compiled binaries are available at http://bibiserv.techfak.uni-bielefeld.de/rnashapes/. For Microsoft Windows a graphical user interface and structure graph output are also included.

RNAshapes decreased time (50-100 times faster than previous implementations in Haskell programming language by Sankoff 1985) and space requirements by reimplementing computation of a small set of representative structures of different shapes, computation of accumulated shape probabilities and comparative prediction of

consensus structures approaches in C programming language (Giegerich et al. 2004, Reeder and Giegerich 2005, Steffen et al. 2006).

An RNA shape is an abstract representation of an RNA secondary structure which is inspired by the dot-bracket representation proposed in the Vienna RNA package (Hofacker et al. 1994). An example of folding via RNAshapes program and shape representations are seen in Figure 2.5.

In the ouput area the actual sequence, the free energy value of folded structure, the dot-bracket representation of structure and a more abstract square-bracket representation are included. Also graphical representation of structure can be acquired with color and font configurations and can be saved as bmp file image.

In the dot-bracket representation dots stand for unmatched nucleotides and brackets stand for matched nucleotides which correspond to another matched nucleotide on the other side of hairpin with respect to terminal loop which is represented by a bunch of dots in two brackets '(…..)'. Thereby, it is possible to reconstruct a hairpin or stem-loop structure from dot-brackets representation.

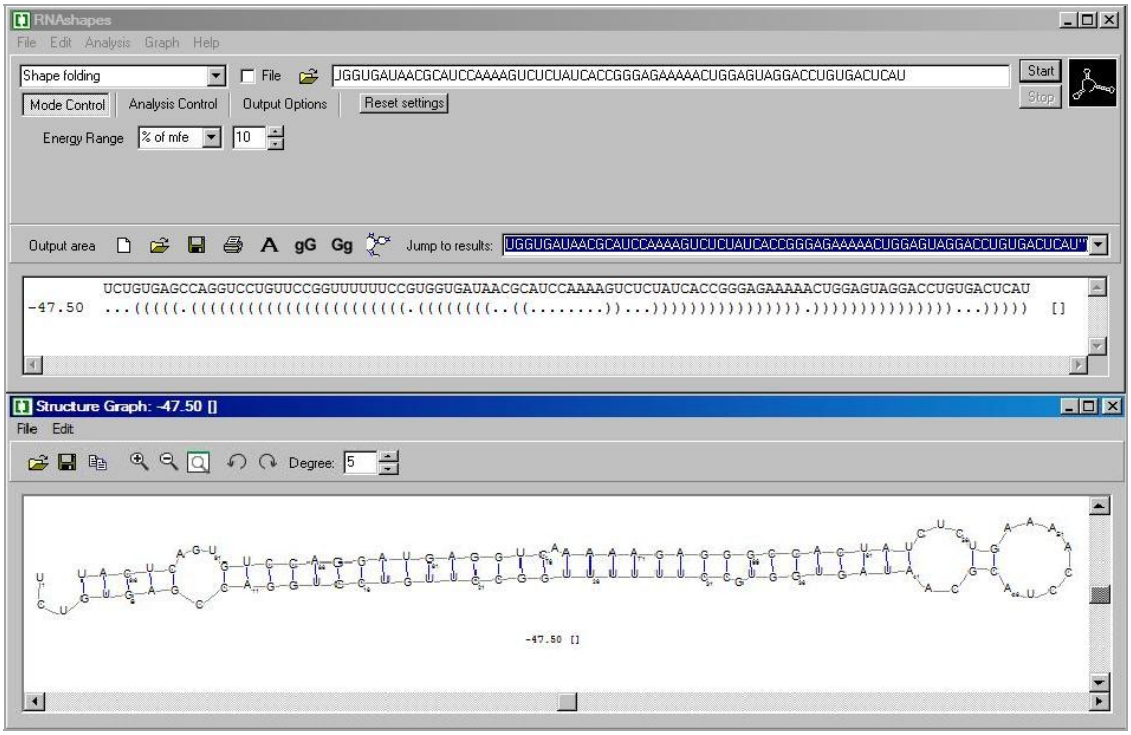

Figure 2.5 Folding of sequences by RNAshapes program. In the output area the are four fields which are the actual sequence(input), minimum free energy field, dot-bracket representation of folded structure and abstract shape. An image of folded structure with nucleotides placed at corresponding positions can be obtained via RNAshapes.

The shapes approach offers five abstraction levels ordered in their degree of abstraction (Figure 2.6) (Steffen et al. 2006). Type one is the most accurate and the most complex abstract shape with the representation of all loops with square brackets and all unpaired regions with underscore characters. Type two consists of representation of loops and unpaired regions in external loop and multiloop but not flanking sequences. Type three has no representation for unpaired regions. Type four represents helix nesting pattern in external loop and multiloop. Type five is the most abstract representation that RNAshapes program exhibits and it offers helix nesting pattern and no unpaired regions.

The square-bracket representation is most abstract output that RNAshapes give. It just represents the distribution of stems, that are evident with matches with a terminal mismatching region, inside a terminal stem. The number of matches or mismatches are not presented in this abstract representation. Thus, a hairpin structure with no loops other than terminal loop ends up with '[]' abstract structure. And more complex structures end up as demonstrated below. However, usually they are not good candidates for pre-miRNAs due to certain reasons; they usually have high minimum free energy (mfe) values thus generate unstable structures, their complex structure can disturb binding proteins, Drosha and Dicer cleavage.

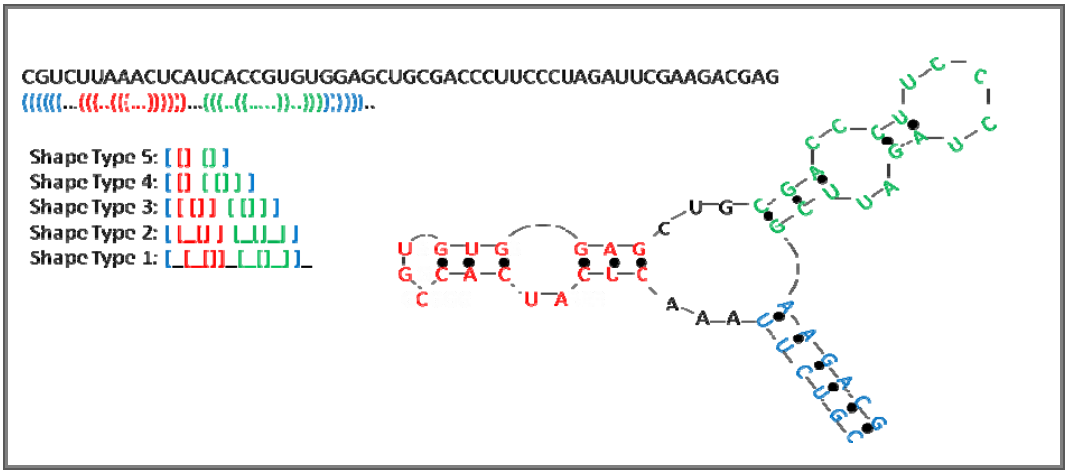

Figure 2.6 Relationship between shape representatives and the folded structure. The colours in dot-bracket representation and abstract shape represents corresponding colours in actualstructure. The structure can be fully regenerated from dot-bracket representation while it is not possible with abstract shapes since they contain a certain level of abstraction and simplification.

Most of the current RNA folding algorithms achieves a bunch results either by calculating a minimum free energy prediction, or a great number of potential suboptimal structures, most of which are redundant therefore expensive in the needs of space and time. Current algorithms are collected, considered and classified in the work of Gardner and Giegerich (Gardner and Giegerich 2004). RNAshapes program uses shape representatives (shreps) which is the structure with the minimum free energy inside a shape class (Steffen et al. 2006). By this way it minimizes the space of consideration.

The RNAshapes package offers a number of functions; input sequences can either be single sequences, sequence files or multi-sequence files in fasta format. Graphical output of secondary structures in postscript format (Hofacker et al. 1994), complete suboptimal folding, detailed options to modulate the program output, extensive command line options, a graphical user interface and interactive visualization of structures. It also introduces a web interface for an online version.

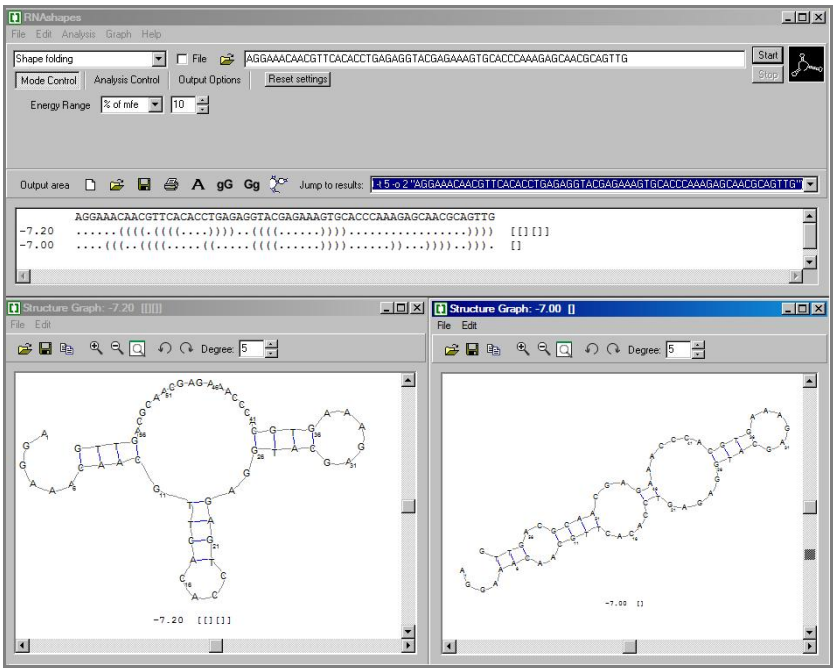

Figure 2.7 A single sequence can lead multiple results with different structures and minimum free energies. Two different structures of a single sequence is shown under output window.

Eventually RNAshapes is a good candidate for our purposes: folding pri-miRNAs to pre-miRNAs, calculating minimum free energies, obtaining good stable hairpins for further consideration.

2.1.4.2. RNAhybrid

RNAhybrid is a tool for finding the minimum free energy hybridisation of two RNA sequences. The tool is primarily meant as a program for microRNA:target duplex evaluation, but certainly it is a handy tool for us to consider the hybridisation states of the ends of Dicer products thus indicating less stable end and decide which strand of final *miRNA:miRNA duplex will be degraded which one will be assembled into RISC in aforementioned manner (Khvorova et al. 2003).

Instead of a single sequence that is folded back onto itself to generate a hairpin structure in the energetically most favourable fashion as RNAshapes, RNAhybrid determines the most favourable hybridization region between two RNA sequences (Rehmsmeier et al. 2004).

Multi-loops, considered in RNAshapes program, are not considered by RNAhybrid (Kruger and Rehmsmeier 2006). These bifurcations are not a part of *miRNA:miRNA duplex thus this property of RNAhybrid rather than disturb our analysis, it expedites our process.

RNAhybrid offers a bunch of options for running from command line. Helix constraint option, maximum length of target sequence, maximum length of query sequence, maximum size of internal loops, maximum lengths of bulge loops etc… Since we want two RNA sequences to pair freely and do not follow up any constraint we did not use any switch of RNAhybrid (Figure 2.8).

```
C:\RNAhybrid>RNAhybrid.exe -s 3utr_fly CAGCAGUA UACAGCAG
target: command_line
length: 8
miRNA : command_line
length: 8

mfe: -8.4 kcal/mol
p-value: 0.000004

position  2
target 5'  A  A    3'
           GC GUA
           CG CAU
miRNA  3' GA  A    5'


C:\RNAhybrid>_
```

Figure 2.8 An example of minimum free energy assessment via RNAhybrid program.

Instead of selecting one strand of mature miRNA duplex, both strands considered as antisense strand and calculated mfe values of ends included in our database. Both strands can be incorporated into RISC with different probabilities due to their mfe values. Thus it is reasonable to investigate potential interactions of both strands. Also it is possible to filtrate undesirable strands out in database easily. So strand selection is accessible at any time with a proper database structure.

2.1.4.3. BLAST

Basic Local Alignment Search Tool (BLAST) is one of the most widely used bioinformatic algorithm for comparing biological sequences which are DNA sequences composed of nucleotides and protein sequences composed of amino acids. Sequence similarity methods can be classified as either global or local. Global similarity algorithms align two sequences by optimizing the alignment of two entire sequences, which often includes large regions of low similarity (Needleman and Wunsch 1970). Local similarity algorithms seek for local similar subsequences, which are local alignments, so a single comparison of two sequences often contains several distinct subsequence alignments while dissimilar subsequence alignments do not considered in the similarity measurement (Goad and Kanehisa 1982, Smith and Waterman 1981). Reasonably local alignment methods are the essence of database searches.

Basically BLAST offers the ability to align a query sequence with a database of sequences, and to identify certain sequences of database that reflect a desired local similarity to the query sequence above a designated threshold. Growing size of sequence databases by genome projects and experimental sequence data emphasizes the need for fast and effective tools for similarity search and integrant data mining methods.

In our study blastn (nucleotide blast) is used to map miRNAs to the genome of *T. gondii*.

2.2. Methods

2.2.1. Initiation

Overall system starts with the parsing of genomic files, both fasta and gff files, of *T. gondii*. Genomic gff file of *T. gondii*, obtained from ToxoDB, contains chromosome names, sequence feature names, start indexes, end indexes, strand information and primary keys of corresponding sequences on first, third, fourth, fifth,seventh and ninth columns respectively of each and every line. Chromosome name designates the chromosome on which the certain sequence located, sequence feature name designates the type of sequence, start indexes are integers and they designate the distance of the starting nucleotide of a sequence to the 5'end of the chromosome. The distance is the number of nucleotides. Similarly end indexes are integers that designate the distance of the ending nucleotide of a sequence to the 5'end of the chromosome. Primary keys are unique identifiers of sequences.

Genomic fasta file of *T. gondii*, also obtained from ToxoDB, contains x number of identifier lines sealed by the greater than character at the beginning of the line '>' which is the trademark of fasta files. The genomic file is departmentalized by the chromosome names of *T. gondii*: each identifier line indicates the sequence of genomic section with the name of chromosome it contains.

Gff and fasta related actions are handled in four Java classes, namely GFF, GFFElements, FASTA, FASTAElements.

GFFElements class just holds variables for chromosome name, sequence feature, start index, end index, strand information and primary keys.

GFF class holds a variable for the path of gff file, one method for parsing gff file and collecting parsed information in corresponding hashmap and one method for sorting hashmap elements according to their start indexes. Hash maps and hash tables are table-like structures which assign specific keys to values. Values do not need to be numbers, characters or simple variables, they can be structured objects as well. Hashmap structures are like hotels with numbered doors and rooms: each key opens one and only one door, but the door does not necessariliy open to only one room. In our particular concern specific keys are primary keys of sequences and values are objects that belongs

to GFFElements class. There are three distinct hashmaps for genes, exons, introns and non-coding regions.

FASTAElements class holds variables for fasta identifier lines, start and end pointers of its corresponding sequence in fasta file (Figure 2.9).

FASTA class holds a variable for the path of fasta file, one method for parsing fasta file and collecting parsed information(identifier lines and indexes of sequences) in corresponding hashmap.

GFF and FASTA classes are connected to each other in another class called GENOMEMap. GENOMEMap class holds a character variable, a FASTA object, a GFF object and a method for getting a desired section of genomic sequence given the GENOMEMap object. The method returns specified length subsequence of an exon, intron or non-coding region sequence. There is three switches to control this initial operation: sequence feature switch, returned sequence length switch, slide length switch. Character variable of GENOMEMap class is used to indicate sequence feature via deciding which hashmap of the GFF object will be used in the process.Sequence length switch decides the length of returned sequence by the getSequence method of GENOMEMap class and this returned subsequence initiates our actual process. Slide length switch decides the number of nucleotides between each subsequence which are subject to consideration. For instance, a slide length of 1 results with consideration of each and every subsequence of sequence. A sequence length of 80 and slide length of 2 is used in our process.

Once the sequence obtained RNAShapes class sets out on (Figure 2.9).

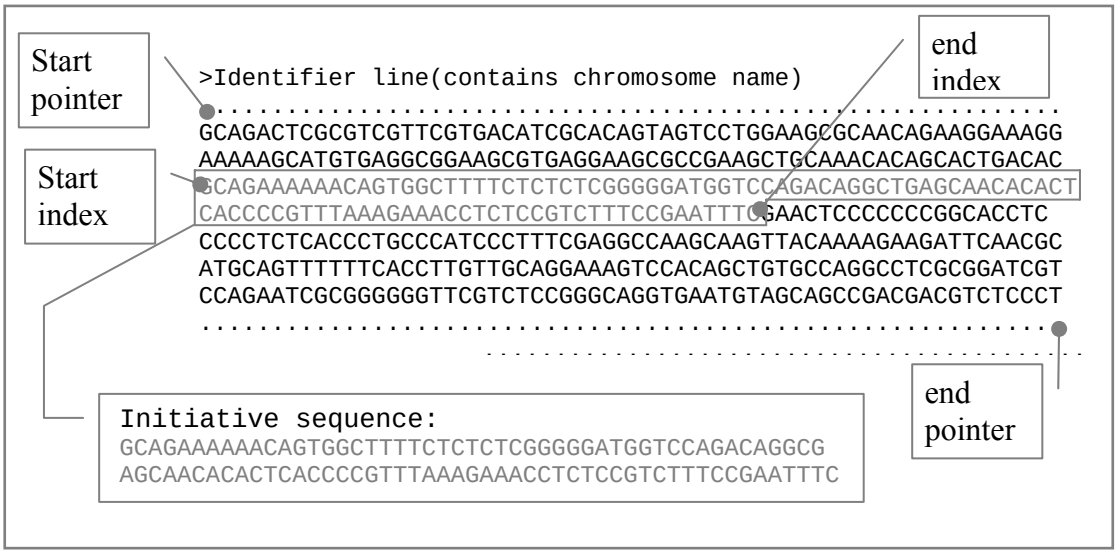

Figure 2.9 The system starts with a section of nucleotides which is 80 nucleotides in length. Start and end indexes in gff file denote the offset of sequence in relevant chromosome. Start and end pointers, which are harvested from fasta file while parsing, denotes the offset of chromosome sequences in fasta file. Sequences are collected given the indexes and pointers. Then each and every 80 nucleotides part of sequences are considered as a pri-miRNA source.

2.2.2. RNAshapes and Folding to Hairpins

RNAshapes class basically and utterly handles folding of sequences: folds sequences, filters folded structures due to some constraints and collect appropriate ones.

RNAshapes class holds reverse-complement of initiative sequence in a field called source. It also has a method to call RNAshapes program to fold sequences, another method to read output of RNAshapes program, a subclass called RNAshapesSubClass to collect the information of RNAshapes program output.

RNAshapesSubClass class has tree fields - one field holds the minimum free energy calculation value regarding a folded RNA, one field holds the shape of folded RNA and other field holds the abstract shape of a folded RNA- and a filtering methods collection. All initiated RNAshapeSubClass objects are subject to filtration.

One constraint in filtration is the minimum free energy (mfe). If the minimum free energy value for a particular structure is higher than a certain threshold it is discarded. The threshold is assessed via folding all miRNA hairpins, which are ±80 nucleotides in length, in miRBase: the microRNA database. The mean and standart deviations of miRBase mfe assessment are -38.045 and 8.214 kcal/mole respectively.

Another filtration step filters out abstract shapes of folded hairpins. Since additional loops consume some number of nucleotides, having lateral loops increases the length of pri-miRNA which can effectively produce *miRNA:miRNA duplex. Perhaps this fact decreases radius of action and disturbance of proteins/enzymes - that take role in miRNA biogenesis – by these additional loops. Therefore this filtration step discards pri-miRNAs which have abstact shapes other than "[]".

Other steps filtrates shape representations of pri-miRNAs due to the length of longest stretch(the length of stretch from SD junction to terminal loop), mismatches allowed in the longest stretch, terminal loop length, length of the arms of shape and mismatches allowed in these arms. Although terminal loop length is not directly influential in the miRNA biogenesis in some cases it slightly affects the process (Han et al. 2006). The length of shape, its longest stretch, mismatches allowed in th shape and mismatches allowed in the longest stretch are investigated since the length of pre-miRNA and mismatches are crucial in the process of miRNA thus values of 25, 20, 8 and 3 are used respectively in our process. Small number of mismatches in a hairpin structure causes a stable and flat structure while a vast number of mismatches, bulges and loops unstabilize the structure.

After folding and filtering, RNAshapesSubClass objects that succeed filters are collected in a vector and sent to Drosha cleavage.

2.2.3. Drosha Cleavage

Another class called Drosha is responsible for whole Drosha cut related operations. The initiative method of Drosha process takes RNAshapes objects as parameter and collect sequences that result from Drosha cut in the vector of Drosha class.

Two arms of hairpins have unsynchronized shape representations due to different number of mismatches, bulges or loops on these arms. The master method of Drosha class synchronizes shape representations of two hairpin arms by introducing gaps where needed. This is a mandatory step owing to exingency of cleavage from accurate locations on both arms.

Another crucial step, being accomplished by a corresponding method of Drosha class, is pinpointing SD(single stranded RNA : double stranded RNA) junction since the

cleavage location of Drosha enzyme is denotes a certain distance from SD junction which is controlled by a switch in the process. The distance is specified as 10 nucleotides from SD junction in coherence with the study of (Han et al. 2006). To pinpoint the SD junction lower stem matches of hairpin is located. The minimum length of the matching stretch to be considered as SD junction is controlled by another switch. The length is set to be 2 nucleotides in our system.

After SD junction is located the cleavage method of Drosha class cleaves hairpins and collects resultant pre-miRNAs in a vector.

2.2.4. Dicer Cleavage

Dicer class has a similar structure with Drosha class. It has a master method to initiate the process and a vector to collect resultant sequences which are RNA duplexes. The lengths of RNA duplexes (Dicer product length) are controlled with a switch, different lengths of products can be obtained if desired. In our study a product length of 22 nucleotides is used in consistence with plenty of studies (Ji 2008, Ketting et al. 2001, Knight and Bass 2001, MacRae et al. 2007).

Dicer class hires two more switches; one switch to control the number of bulges in the product of Dicer, other is to control the minimum length of the group of mutual mismatches to be considered as a bulge. In consistence with these constraints Dicer class cuts and filters Drosha products then collects Dicer products in a vector for further processing and filtering.

2.2.5. RNAhybrid and Strand Selection

RNAhybrid program is hired to implement the strand selection to be able to decide which strand of RNA duplex(Dicer product) will be considered as antisense strand and which strand of duplex will be considered as guide strand.

Strand selection is carried out in consistence with the framework proposed and established firmly in several pioneering studies (Khvorova et al. 2003, Schwarz et al. 2003, Tomari et al. 2007).

RNAhybrid class has only one method to format strands and a subclass called sectionsStrands to collect corresponding information of formatted strands. In this

method strands are divided into five parts in order to calculate relative thermodynamical stabilities. The parts are 5’ overhanging sequence (2 nucleotides long), 5’ end of sequence, middle part of sequence, 3’ end of sequence and 3’ overhanging sequence (2 nucleotides long) respectively. After formatting these sequence parts passed to the fields of a sectionsStrands object and invoke its methods.

SectionsStrands class has aforementioned fields and two methods; one method invokes RNAhybrid program for both 5’ and 3’ ends and reads the output, other method compares the mfe values of these sides. The relative flexibility on the 5’ end causes preferred unwinding and provokes selection of certain strand.

After strand selection there is a step that conducts checking the length of final antisense miRNA length. Afterall the length of antisense miRNA can be improper after annihilation of gaps that introduced to synchronize sequences. If antisense miRNA length is shorter than a desired threshold it can be discarded here. The length it is used in our process is the same as the Dicer product length which is 22.

Subsequent to strand selection RNAhybrid class delivers selected strand to another class called BLAST where BLAST gets significant hits on genome.

CHAPTER 3

RESULTS AND DISCUSSION

The aim of our study is to identify all potential miRNA sources and all their potential interaction targets. Interactions are presented in a relational database with tables containing all relevant values for further filtering, distinction and for presentation and retrieval purposes.

Sequences are presented in the table "sequences" with their related features: type of sequence, query start offset, free energy value, lengths of calculated RNA shape, number of mismatches in the RNA shape, length of longest stretche in the RNA shape and its number of miss matches, number of lower stem matches which denotes the SD junction, length of the terminal loop, free energy value of terminal ends, antisense sequence, sequence of query in BLAST alignments, start and end indexes of BLAST query in antisense strand, start and end indexes of the database hit, feature of potential source sequence, chromosome on which the potential source is located, sequence and shape of pri-miRNA sequence, strand on which sources are located, mono-, di- and tri-nucleotide counts of antisense sequence and the lengths of the flanking ends. The type of sequence can either be source or target. Offset is the distance of the 80 nucleotide source sequence from the beginning of the corresponding exon, intron or non-coding region. The free energy value is calculated for hairpins by the RNAshapes program. Pri-miRNA sequence, length of its RNA shape (dot-bracket representation), number of mismatches in the RNA shape, the length of the longest stretch in the RNA shape, the number of mismatches in the RNA shape, the number of lower stem matches and the length of the terminal loop are calculated from folded RNA structures which are produced by the RNAshapes program (Figure 3.1). The free energy values of the terminal ends are calculated by the RNAhybrid program (cf. Materials and Methods). Antisense sequences are sequences obtained after sequential endonucleolytic cleavages and pertinent filtering steps. BLAST related values are obtained after BLAST queries of antisense strands. Source and target features are received from GFF file of *T. gondii* with the methods of the GFF java class developed in this study.

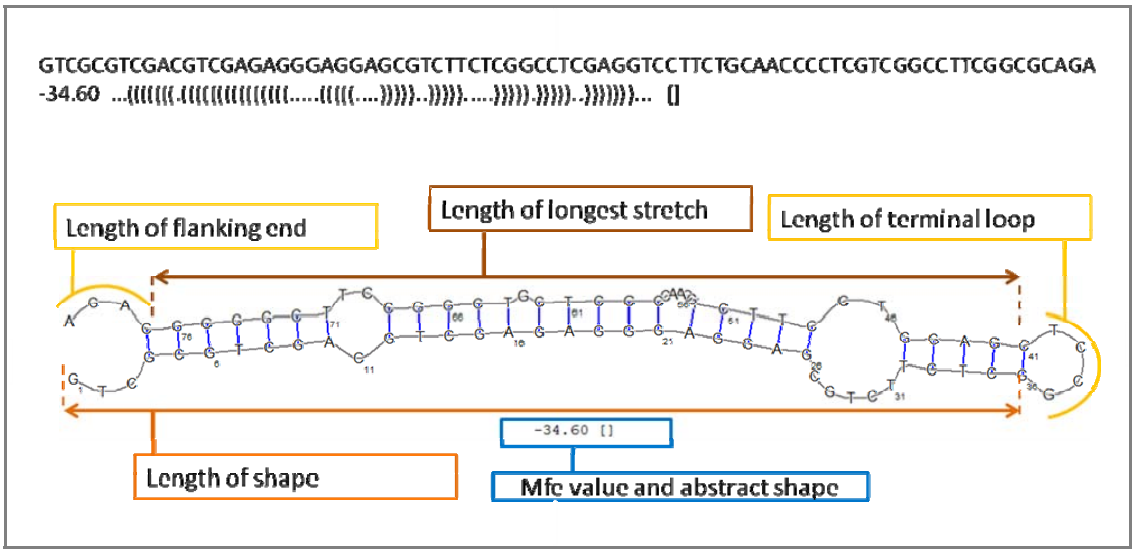

Figure 3.1 Shape-related values which are used to evaluate a folded pri-miRNA. Each shape has two arms connected to terminal loop (These arms are referred as "top" and "bottom" in our database). Length of the shape is the distance between terminal loop and one terminal end. Length of the longest stretch is the distance between terminal loop and SD junction. Terminal loop is the group of central circular mismatches. Number of mismatches and matches in these regions are subject to consideration with different emphasis.

The significance of the hairpins identified in this study are evaluated by the values obtained from previously, experimentally identified hairpin sequences. The hairpins sequences are obtained from the file hairpin.fa from miRBase (Ambros et al. 2003, Griffiths-Jones 2004, Griffiths-Jones et al. 2006, Griffiths-Jones et al. 2008, miRBase 2009). For a number of measurable properties mean and standart deviation values have been calculated which are used to set the thresholds in our system (Appendix A). The values are calculated for sequences which are greater than 75 nucleotides and lesser than 80 nucleotides in length.

Interactions are presented in the "interference" table with interaction related features which are the source sequence id, the target sequence id, the length of the alignment between source and target sequences, the e-value of the alignment, the number of matches, mismatches and gaps in the alignment.

The interactions are evaluted with pertinent sql queries like the one shown below:

```
SELECT *
FROM Sequences AS src INNER JOIN Interference AS i ON
src.ID=i.srcSeqID
```

```
INNER JOIN Sequences AS tar ON tar.ID=i.tarSeqID
where
    src.rsmfe < -35 and
    src.di > 11 and
    eVal < 0.01 and
    src.tri > 17 and
    src.lsmlength > 3 and
    algnlength > 19 and
    tar.seqFeature='outer' and
    src.shpLength > 33 and
    src.loopLength < 10 and
    src.shpMM > 2 and
    src.shpMM < 12
```

It was possible to identify a large number of interactions in the *T. gondii* genome with loose thresholds. More restrictive filtering may lead to more significant results but it may also restrict the usability of the created database since some interactsions may be lost due to filtering. Therefore the filtering criteria were not too restrictive and the significance of a given interaction can be calculated from the values presented in the database.

From the large pool of potential RNAi interactions a few shall be considered in the following. One of the potential miRNA sources is the TGME49_000300 gene. The actual source is at the position 307, the third intron of the TGME49_000300 gene, on the minus strand (Figure 3.2). TGME49_000300 is located on the 8^{th} chromosome. The miRNA derived from the gene TGME49_000300 has one potential significant target. The pri-miRNA of the TGME49_000300 gene contains four mismatches on one arm and eight mismatches on the other. It contains a six nucleotide terminal loop, -42.5 kcal/mol minimum free energy value and has no flanking nucleotides.

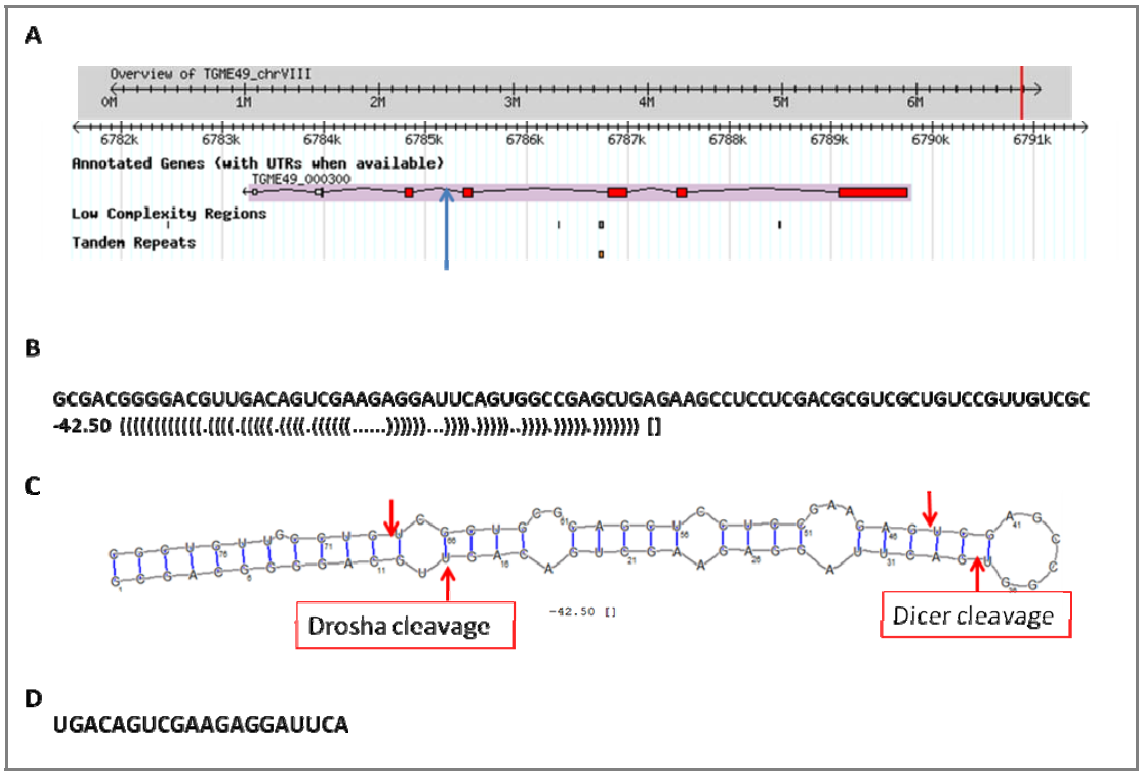

Figure 3.2 Location and folding of the potential miRNA source from the TGME49_000300 gene. Field A shows the gene structure and the location of the gene. Red rectangles denote exons, interconnecting lines denote introns, while rectangles denote low complexity regions and tandem repeats. The blue arrow points to the start of the source of the potential miRNA. Field B shows the pri-miRNA sequence, the dot-bracket representation of the folded pri-miRNA, the free energy value of the RNA shape and the abstract shape representation of the folded pri-miRNA. Field C shows an image of the folded pri-miRNA. Red arrows indicate possible cleavage sites of Drosha and Dicer enzymes. Field D shows the mature miRNA sequence (Source: ToxoDB 2009).

The pri-miRNA and the mature miRNA of the TGME49_000300 gene seem to have a complex sequence compositon and are thus not filtered by our low complexity thresholds. The mature miRNA contains all four nucleotides, 12 different dinucleotides and 18 different trinucleotides. However, low complexity in nucleotide composition is not always a handicap for miRNA sequences. Low complexity tandem repeats of the genome can serve as good sources of miRNA. Tandem repeats give pri-miRNAs a good potential of folding and a low free energy. Nevertheless, the repetitive sequences in pri-miRNAs can lower the specifity of miRNA candidates and increase their potential interactions so much that they would be difficult to handle. However, there are several identified miRNAs with repetitive sequences (Figure 3.3). It is not surprising since a

pri-miRNA can have several repeats while it still retains a certain amount of complexity. We identified a lot of repetitive sequences with a great number of targets. However, it is possible to filter them out through complexity filtering by mono-nucleotide, di-nucleotide and tri-nucleotide counts if low complexity sequences were among the miRNA examples .

I should emphasize that strand selection by free energy difference between terminal ends predicted the selection of the mature strand of TGME49_000300 miRNA correctly.

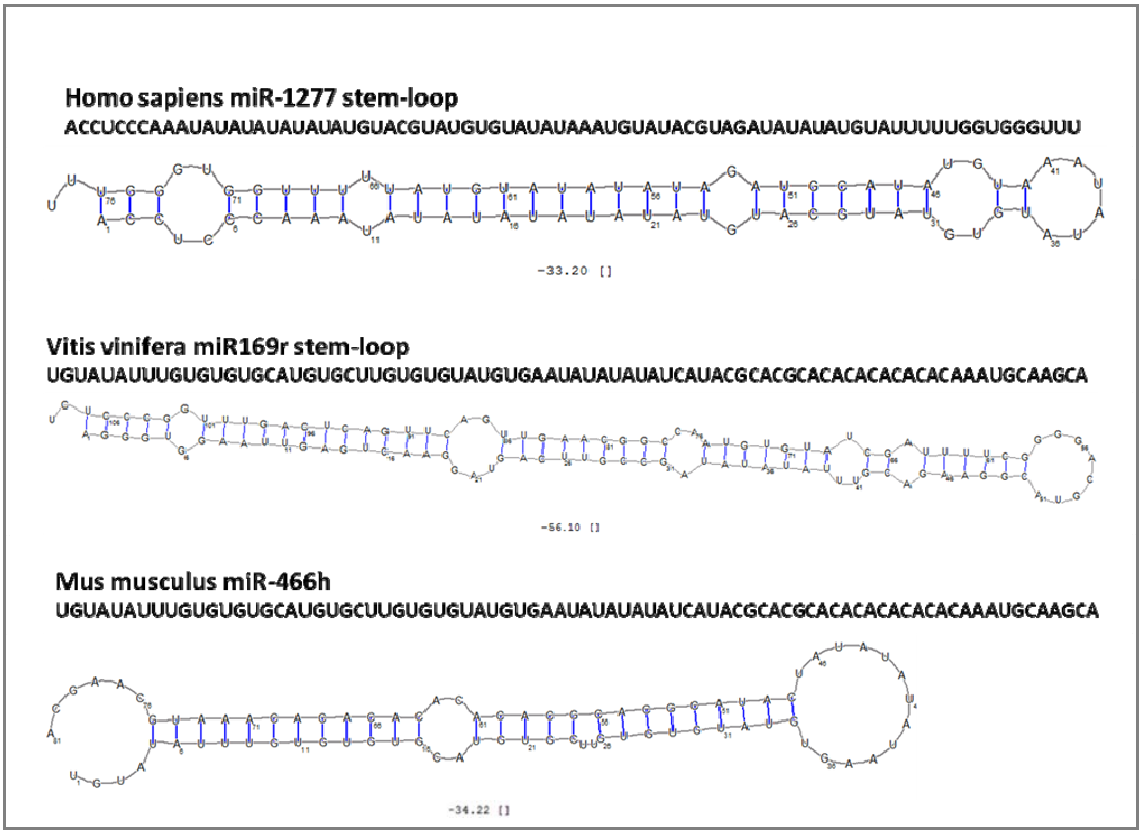

Figure 3.3 Examples of previously identified hairpins with tandem repeats (Source: miRBase 2009).

The potential miRNA from the TGME49_000300 gene interacts with its target with perfect complementarity (Figure 3.4). The TGME49_000300-miRNA interacts with the non-coding region following the hypothetical protein gene TGME49_068320 on the plus strand. The location of the interaction is between positions 6567516 and 6567535 on the eigth chromosome. The index of the gene TGME49_068320 is 6542055 to 6548980. The TGME49_000300-miRNA:target interaction seems to be located far

away from the 3'UTR(untranslated region) of the TGME49_068320 gene thus an interference between these elements does not seem feasible although long range interactions are not surprising given the overall RNAi mechanism in other organisms.

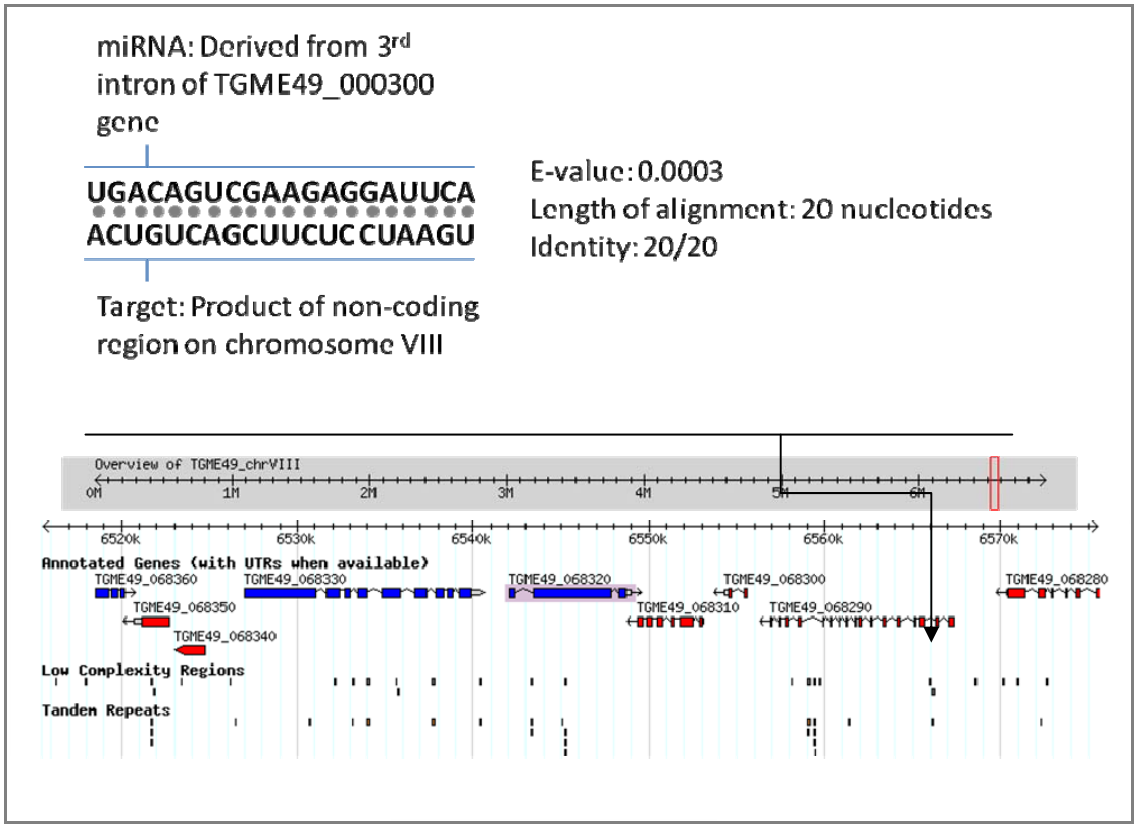

Figure 3.4 Interaction of the TGME49_000300 product miRNA and its target. The image shows the location of the genes around the TGME49_068320 gene. It has been reproduced from ToxoDB (Source: ToxoDB 2009).

Another significant interaction seems to be formed between two potential miRNAs of the TGME49_003990, a hypothetical protein encoding gene and its two potential targets (Figure 3.5). The TGME49_003990 gene is located on chromosome TGME49_chrVIIa on the minus strand between the indeces 2020380 and 2022951. One of the pri-miRNAs derived from the TGME49_003990 gene has four mismatches on one arm and eight mismatches on the other. It also has a four nucleotide long terminal loop and a free energy of -39.1 kcal/mol. The other pri-miRNA derived from the TGME49_003990 gene has two mismatches on one arm and eight mismatches on the other. It has a seven nucleotide long terminal loop and a free energy of -38.4 kcal/mol. Both of the two mature miRNAs of the TGME49_003990 gene are composed of four

different nucleotides, 12 different di-nucleotides and 18 different tri-nucleotides which gives them a complex sequence composition.

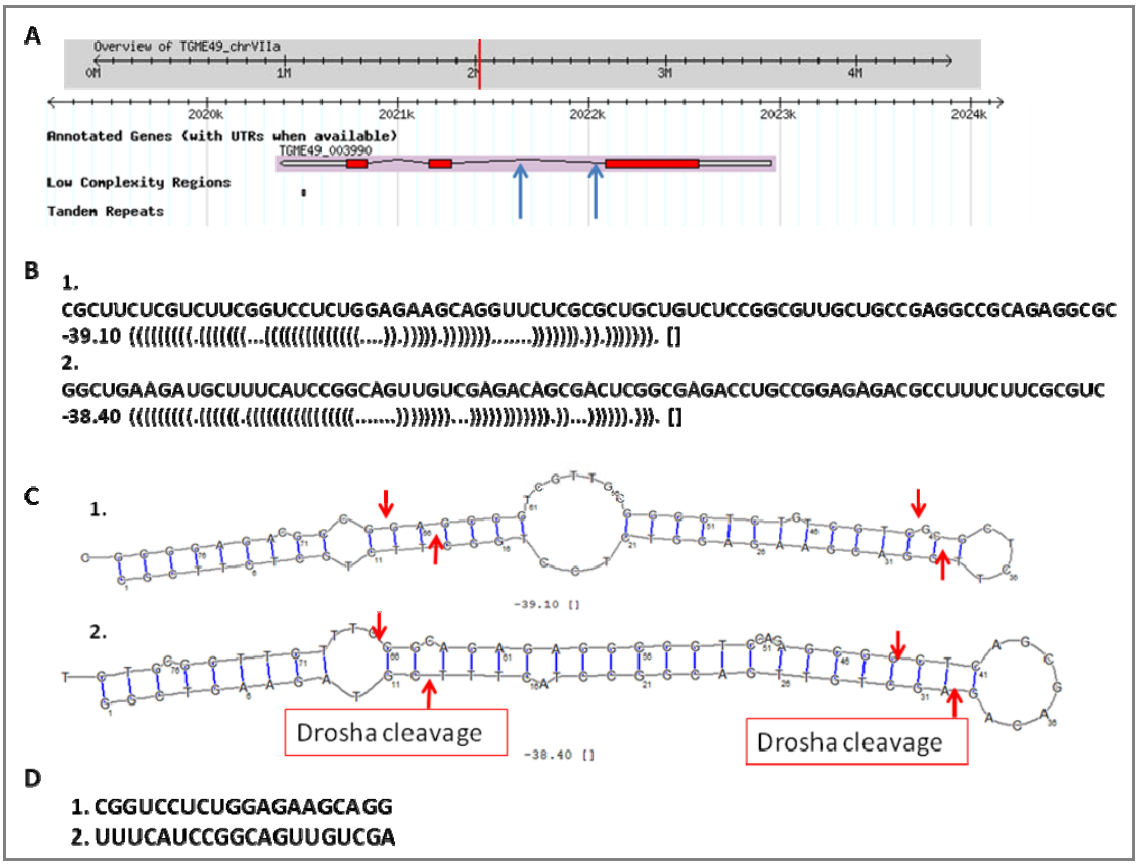

Figure 3.5 Location and folding of potential pri-miRNAs from the source gene TGME49_003990. Field A shows the gene structure and location of the gene. Red rectangles represent exons, interconnecting lines denote introns. The blue arrows points to the sources of the potential miRNAs. Field B shows the pri-miRNA sequences, the dot-bracket representation of the folded pri-miRNAs, the free energy of the RNA shape and the abstract shape representations of the folded pri-miRNAs. Field C shows images of the folded pri-miRNAs. Red arrows indicate potential cleavage sites of Drosha and Dicer enzymes. Field D shows the mature miRNA sequences (Source: ToxoDB 2009).

The two mature miRNAs of the TGME49_003990 gene have two potential interferences with two non-coding regions following the TGME49_004280 and TGME49_032290 genes (Figure 3.6). Both interactions have quite acceptible e-values (0.0003) and sequence identities. However, the 3'UTR regions of the preceding genes seem to lie a considerable distance away from the location of interaction (Figure 3.6).

Nevertheless, without a collection of proper 3'UTR in *T. gondii* it is hard to assess the quality and significance of the interference.

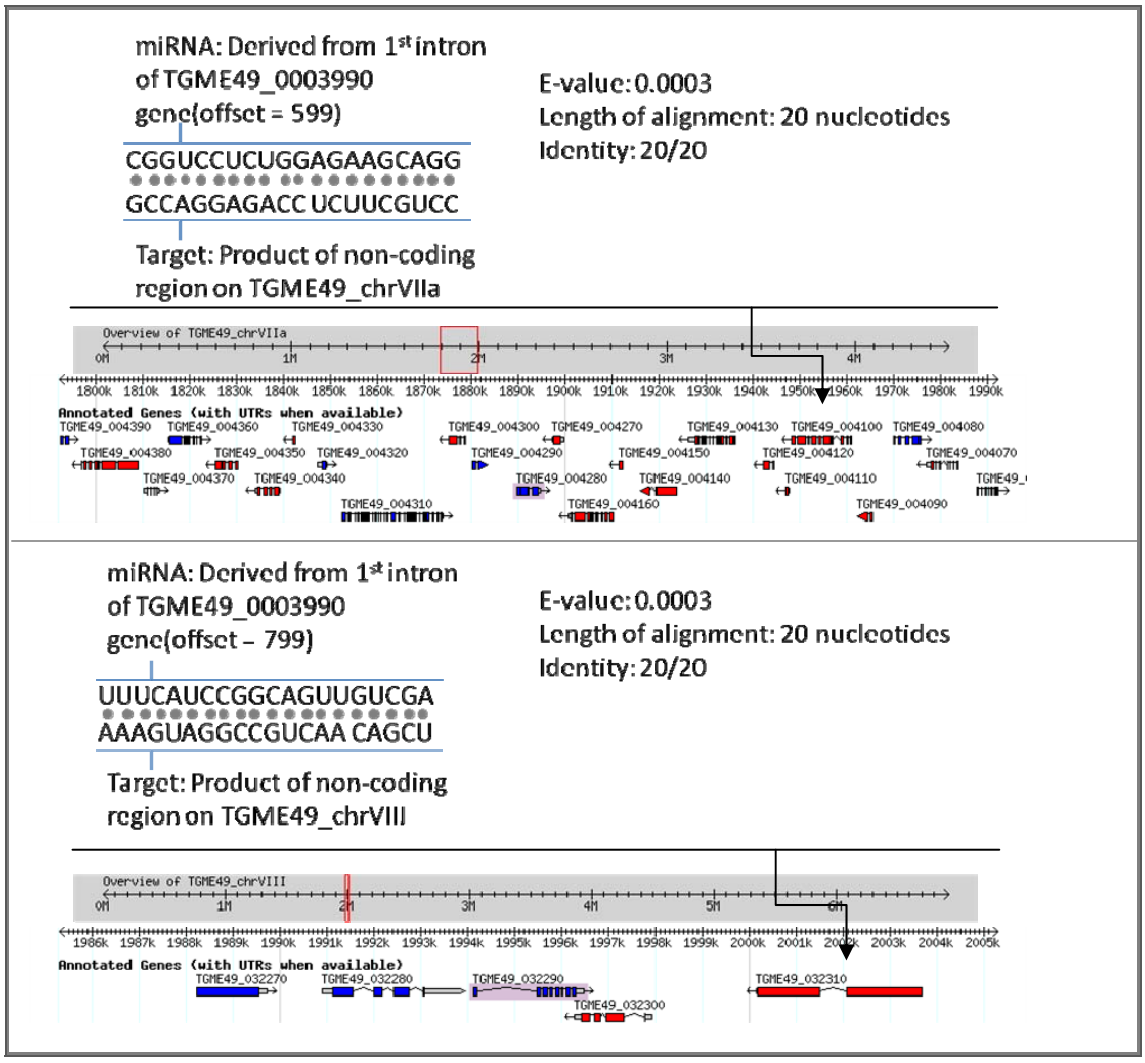

Figure 3.6 Interaction of TGME49_000300 product miRNAs and their targets. Arrows point the location of target non-coding regions. Rectangles on overview of the chromosomes show the location of interaction regions on chromosome. Rectangles oriented by an arrow on one end denotes annotated coding regions (arrows that head towards right denote coding regions on plus strand while arrows that head towards left denote cogind regions on minus strand) (Source: ToxoDB 2009).

An alternative pri-miRNA source lies on the second intron of the TGME49_065140 gene on the chromosome TGME49_chrIX. The TGME49_065140 gene is a hypothetical gene region on the minus strand. The pri-miRNA is derived from TGME49_065140 and has 6 mismatches on each arm, one nucleotide on both terminal flanking ends, a four nucleotide long terminal loop and a free energy of -39.1 kcal/mol.

The mono-, di- and tri- nucleotide counts are 4, 13 and 21 respectively which emphasizes the complexity of the pri-miRNA's sequence composition. The locus of the TGME49_065140 gene does not contain any tandem repeats or signs of low complexity.

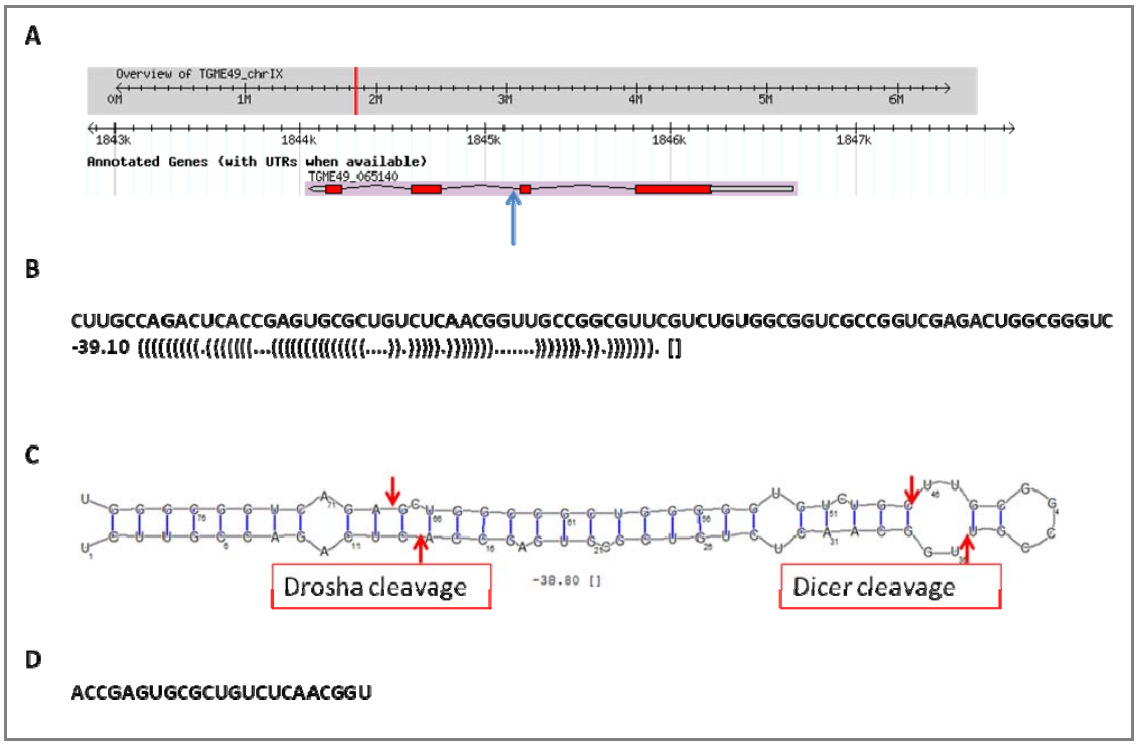

Figure 3.7 The Location and folding of the potential pri-miRNA from the source gene TGME49_065140. Field A shows the gene structure and location of the gene. Red rectangles denote exons, interconnecting lines represent introns. The blue arrow points to the source of the potential miRNA. Field B shows the pri-miRNA sequence, the dot-bracket representation of the folded pri-miRNA, the free energy value of the RNA shape and its abstract representations. Field C shows images of the folded pri-miRNA. Red arrows indicate the potential cleavage sites of Drosha and Dicer enzymes. Field D shows the mature miRNA sequence (Source: ToxoDB 2009).

The interaction between the TGME49_065140 miRNA and its target exposes perfect complementarity with a considerable e-value of 7e-006 (Figure 3.8). The location of interaction (between positions 1785577-1785599) lies between the gene TGME49_065240 and the TGME49_065220 co-chaperone gene.

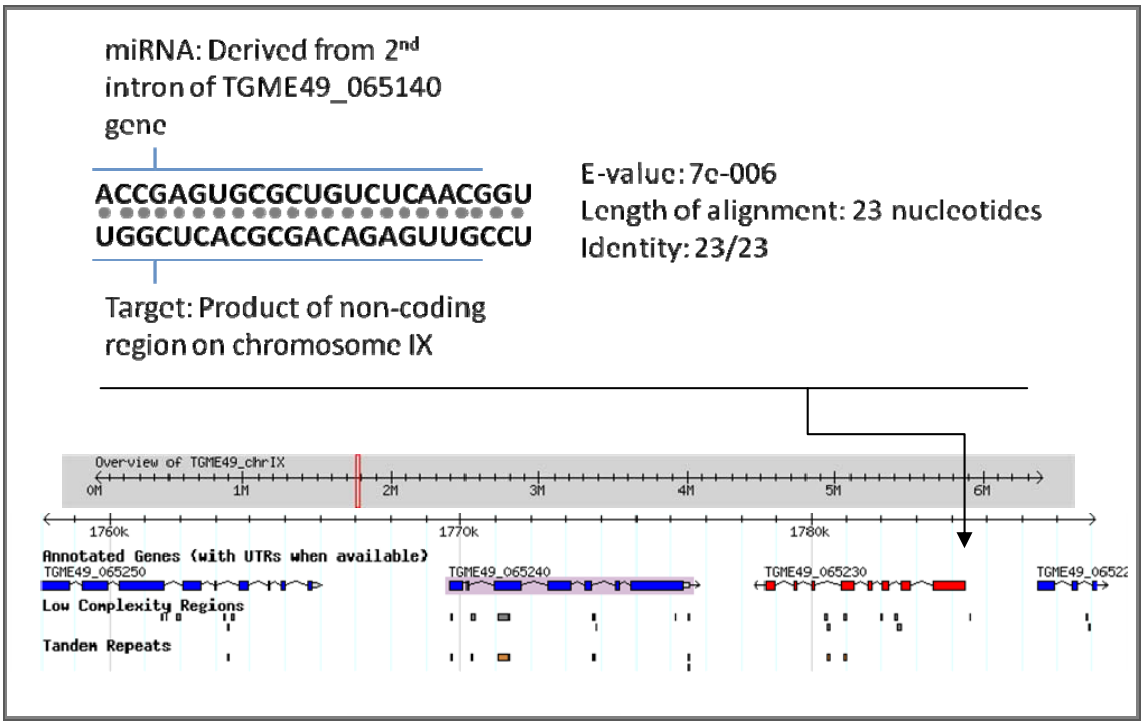

Figure 3.8 Interaction of TGME49_000300 product miRNAs and their targets. Arrow point the location of target non-coding region. Rectangle on overview of the chromosomes show the location of interaction region on chromosome. Rectangles oriented by an arrow on one end denotes annotated coding regions (arrows that head towards right denote coding regions on plus strand while arrows that head towards left denote cogind regions on minus strand) (Source: ToxoDB 2009).

Interactions presented so far are just a few examples around many. With different threshold sets in database queries, different number of sources, targets and interactions can be identified. We have identified collection of interactions and mappings with three threshold sets set1, set2 and set3. Every threshold set contains four delimiters inferred from the properties of miRBase hairpins. Delimiters are mfe value of folded pri-miRNA, di-nucleotide count, tri-nucleotide count and number of mismatches in folded pri-miRNAs.

Threshold set set1 consists of miRNA candidates with minimum free energy interval between -40 kcal/mol and -35 kcal/mol, di-nucleotide count interval between 10 and 12, tri-nucleotide count interval between 15 and 19, number of shape mismatches interval between 6 and 8.

Threshold set set2 consists of miRNA candidates with minimum free energy interval between -35 kcal/mol and -30 kcal/mol, di-nucleotide count interval between 8

and 10, tri-nucleotide count interval between 13 and 15, number of shape mismatches interval between 4 and 6.

Threshold set set3 consists of miRNA candidates with minimum free energy interval between -30 kcal/mol and -25 kcal/mol, di-nucleotide count interval between 6 and 8, tri-nucleotide count lesser than13, number of shape mismatches lesser than 4.

A database query with set 1 ends up with many one-to-one interactions and four more complex interactions (Figure 3.9, 3.10).

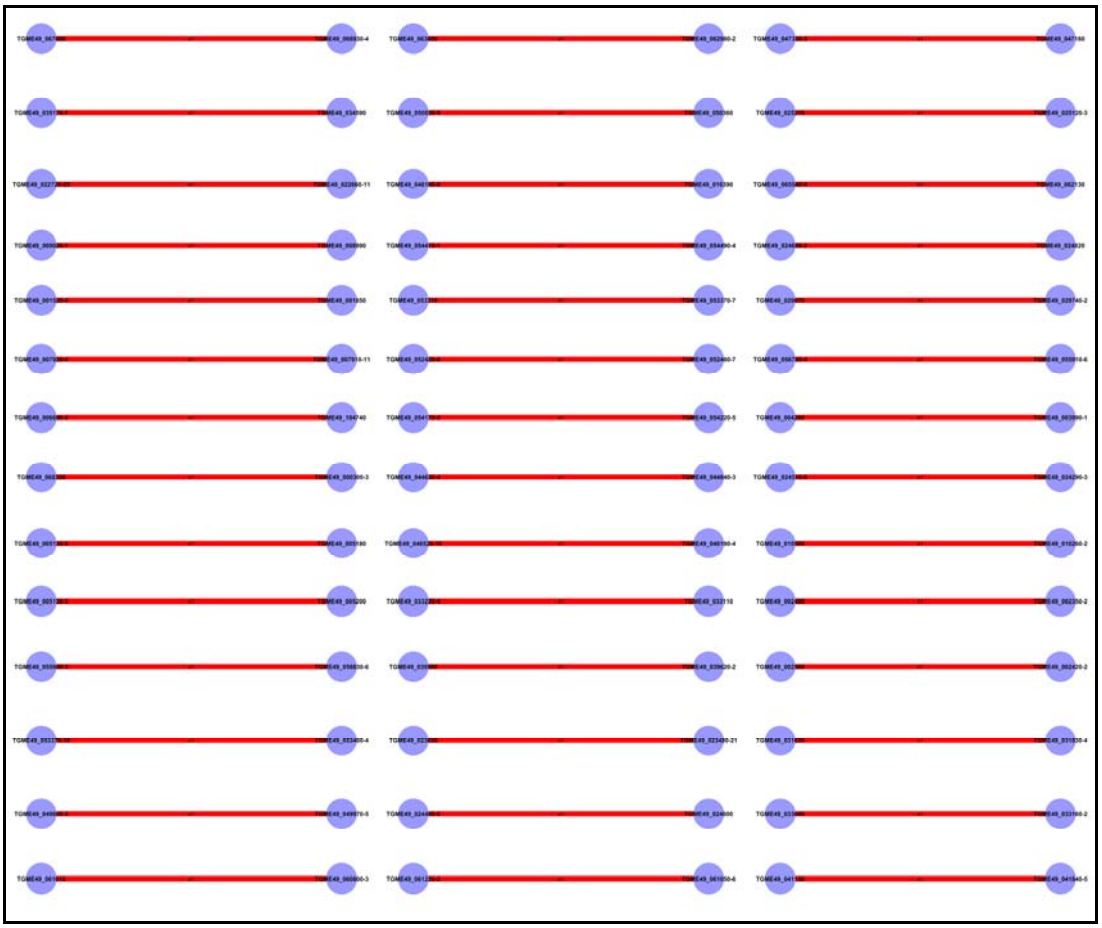

Figure 3.9 One-to-one interactions identified by the set1 threshold set. Nodes denotes sources and targets while edges denotes interactions between connected nodes.

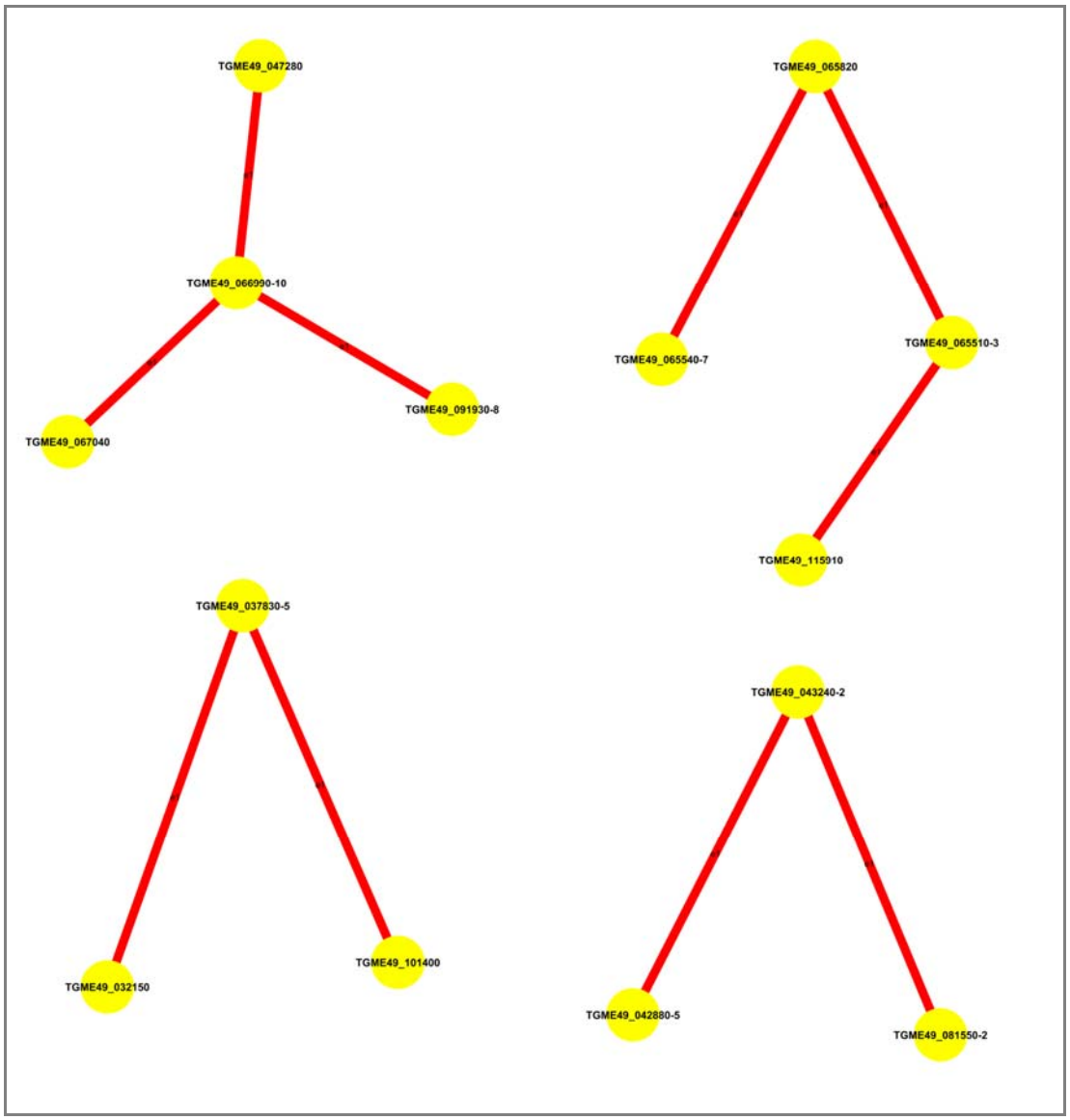

Figure 3.10 Multiple interactions identified by the set1 threshold set. Nodes denotes sources and targets while edges denotes interactions between connected nodes.

Queries with set2 and set3 results in numerous complex interactions. Database mining with different delimiter sets can identify different sources, targets and interactions (Figure 3.11, 3.12).

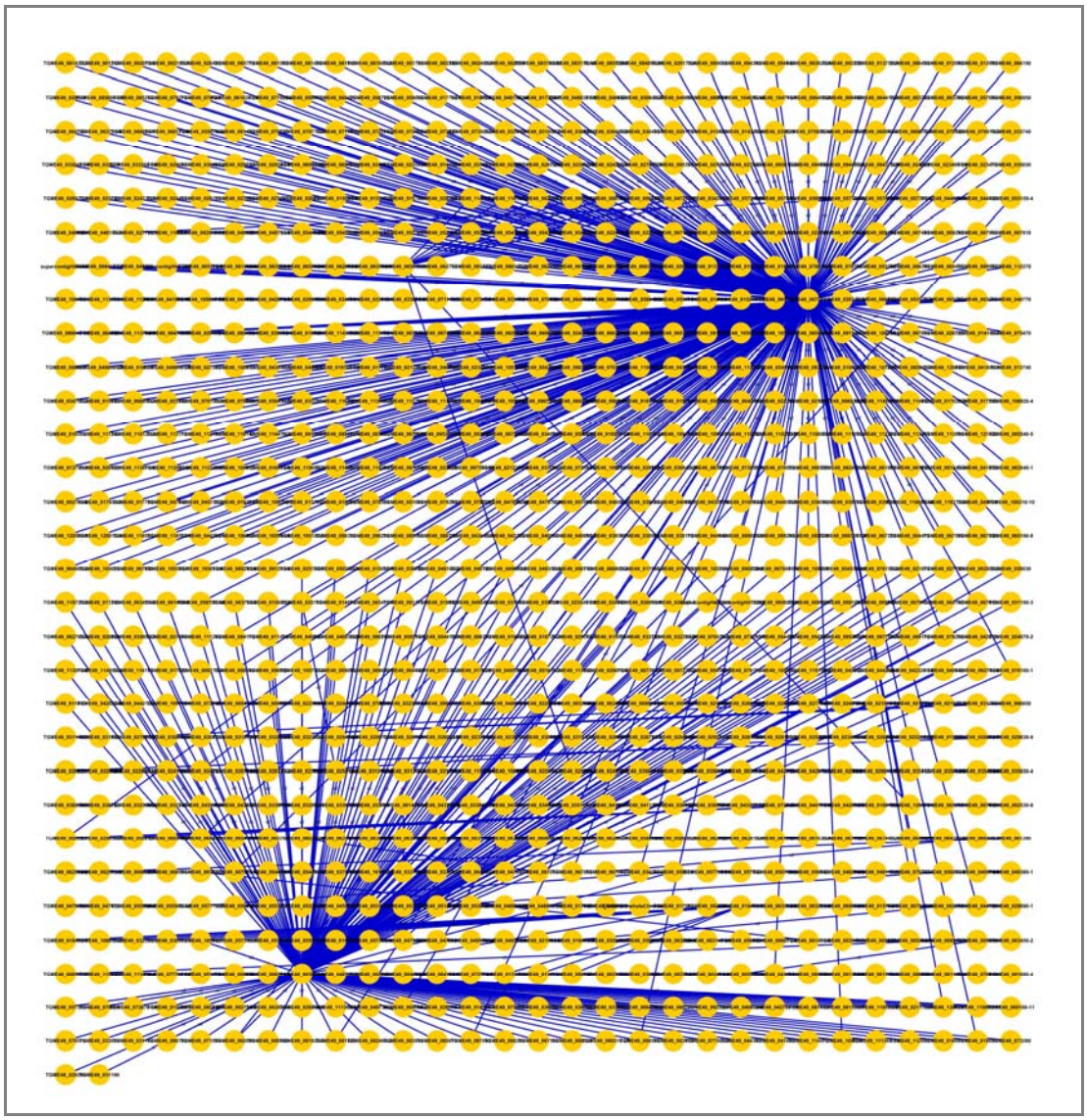

Figure 3.11 Multiple interactions identified by the set2 threshold set. Nodes denotes sources and targets while edges denotes interactions between connected nodes.

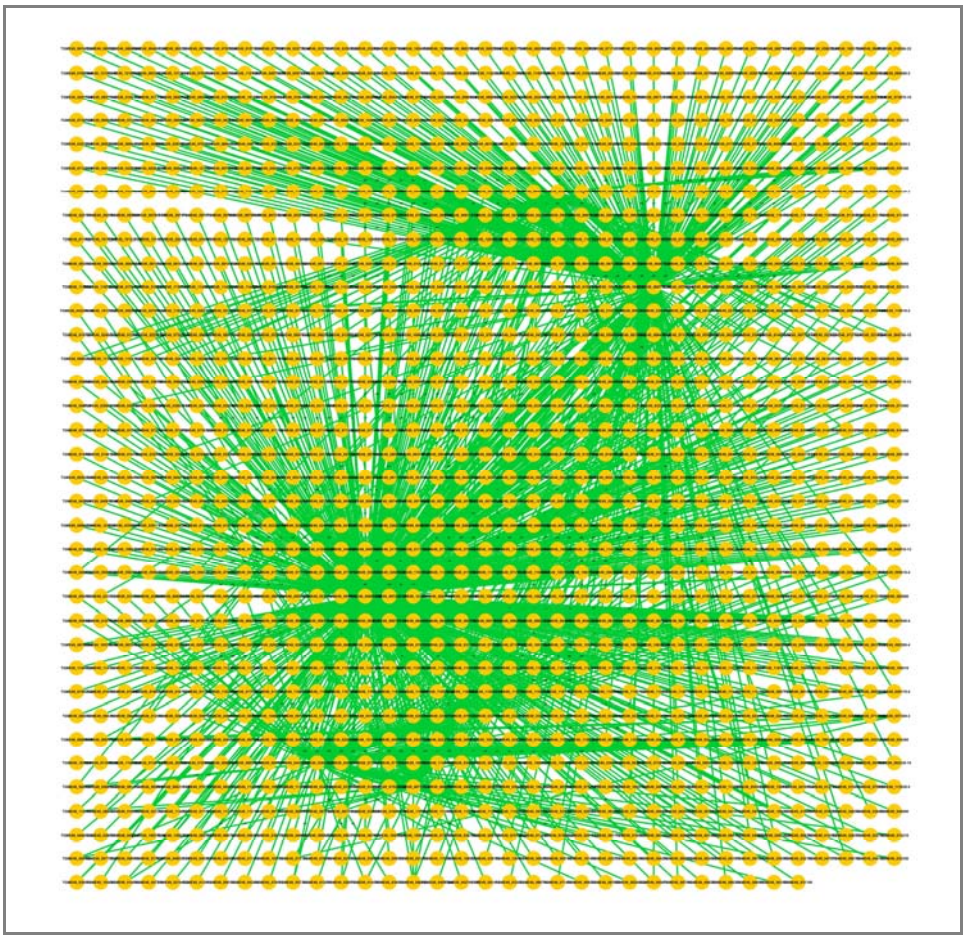

Figure 3.12 Multiple interactions identified by the set3 threshold set. Nodes denotes sources and targets while edges denotes interactions between connected nodes.

It is hard to evaluate potential interferences without the collection of annotations which includes untranslated and non-coding regions. It can be shown with e-values and sequence complexities that identified interactions may exert strong and specific interactions. However, without information about interaction locations the scope and influence of interaction can not be decided properly. An experimental evaluation is always needed to confirm results achieved by genome wide data mining. Clearly, the data represented in the database produced in this study enables the research community to evaluate if unexpected experimental outcome could stem from RNAi interactions. With the data generated here targeted RNAi ineraction analyses can be designed to evaluate some of the ineractions from the database and thus put it on a foundation of experimental data.

CHAPTER 4

CONCLUSION

Recent studies have shown several effects and properties of miRNAs in many organisms. MiRNAs can control gene expressen both in the initiation or post-initiation step of translation. They are known to be very important in development and tissue specific gene expression.

Recent knowledge on RNAi regulation in *T.gondii* is not adequate to disclaim or to validate existence of RNAi regulation in *T.gondii*. However there are promising experimental evidences on possible RNAi functionality in *T.gondii*.

There is no identified Drosha and Dicer homologues in *T.gondii* in the time being. However, there can be different mechanisms that can end up with miRNAs. It has known that some intronic pri-miRNAs (mirtrons) of *D.melanogaster* and *C.elegans* have the ability to by-pass drosha cleavage (Ruby et al. 2007). Besides, the existence of AGO homologue with both Paz and Piwi encourages studies on RNAi regulation in (Ullu et al. 2004).

Our study exhibits the existence of potential pri-miRNA sources and their targets in the genome of *T.gondii*. There are significantly specific interactions between miRNA candidates, which are derived from potential pri-miRNA sources, and their targets. Different number of interactions are obtained with different properties by using several threshold sets. Furthermore, we constructed a database with numerous possible interactions in T.gondii. It is for sure that our database contains many false positives but filtration with desired restrictions can easily be done in database. Thus, it would be beneficial to experimental scientists to use obtained data to aid their experiments on RNAi regulation in *T.gondii*.

The system we implement in Java™ can be used on other organisms as well. Potential RNAi assessment of available genome sequences can be done by fine tuning thresholds we used.

REFERENCES

Ajioka, J. W. 1997. The protozoan phylum Apicomplexa. *Methods* 13 (2):79-80.

Al-Anouti, F., and S. Ananvoranich. 2002. Comparative analysis of antisense RNA, double-stranded RNA, and delta ribozyme-mediated gene regulation in Toxoplasma gondii. *Antisense Nucleic Acid Drug Dev* 12 (4):275-81.

Allen, E., Z. Xie, A. M. Gustafson, and J. C. Carrington. 2005. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* 121 (2):207-21.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* 215 (3):403-10.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* 431 (7006):350-5.

Ambros, V., B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl. 2003. A uniform system for microRNA annotation. *RNA* 9 (3):277-9.

Aravin, A. A., G. J. Hannon, and J. Brennecke. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318 (5851):761-4.

Aspinall, T. V., D. H. Joynson, E. Guy, J. E. Hyde, and P. F. Sims. 2002. The Molecular Basis of Sulfonamide Resistance in Toxoplasma gondii and Implications for the Clinical Management of Toxoplasmosis. *The Journal of Infectious Diseases* 185 (11):1637-1643.

Aurrecoechea, C., M. Heiges, H. Wang, Z. Wang, S. Fischer, P. Rhodes, J. Miller, E. Kraemer, C. J. Stoeckert, Jr., D. S. Roos, and J. C. Kissinger. 2007. ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res* 35 (Database issue):D427-30.

Baatz, Holger, Alireza Mirshahi, Joachim Puchta, Hermann Gumbel, and Lars-Olof Hattenbach. 2006. Reactivation of Toxoplasma Retinochoroiditis Under Atovaquone Therapy in an Immunocompetent Patient. *Ocular Immunology and Inflammation* 14 (3):185-187.

Bagga, S., J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus, and A. E. Pasquinelli. 2005. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* 122 (4):553-63.

Bass, B. L. 2000. Double-stranded RNA as a template for gene silencing. *Cell* 101 (3):235-8.

Bastin, P., K. Ellis, L. Kohl, and K. Gull. 2000. Flagellum ontogeny in trypanosomes studied via an inherited and regulated RNA interference system. *J Cell Sci* 113 (Pt 18):3321-8.

Basu, U., K. Si, J. R. Warner, and U. Maitra. 2001. The Saccharomyces cerevisiae TIF6 gene encoding translation initiation factor 6 is required for 60S ribosomal subunit biogenesis. *Mol Cell Biol* 21 (5):1453-62.

Baulcombe, D. 2004. RNA silencing in plants. *Nature* 431 (7006):356-63.

Behm-Ansmant, I., J. Rehwinkel, T. Doerks, A. Stark, P. Bork, and E. Izaurralde. 2006. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev* 20 (14):1885-98.

Berkhout, B., and K. T. Jeang. 2007. RISCy business: MicroRNAs, pathogenesis, and viruses. *J Biol Chem* 282 (37):26641-5.

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

Blader, I. J., and J. P. Saeij. 2009. Communication between Toxoplasma gondii and its host: impact on parasite growth, development, immune evasion, and virulence. *APMIS* 117 (5-6):458-76.

Blaszczyk, J., J. E. Tropea, M. Bubunenko, K. M. Routzahn, D. S. Waugh, D. L. Court, and X. Ji. 2001. Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. *Structure* 9 (12):1225-36.

Bobrow, D. G., and T. Winograd. 1977. An Overview of KRL, a Knowledge Representation Language. *Cognitive Science: A Multidisciplinary Journal* 1 (1):3 - 46.

Bohnsack, Markus T., Kevin Czaplinski, and Dirk Gorlich. 2004. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* 10 (2):185-191.

Borsani, O., J. Zhu, P. E. Verslues, R. Sunkar, and J. K. Zhu. 2005. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* 123 (7):1279-91.

Boyle, J. P., J. P. Saeij, M. D. Cleary, and J. C. Boothroyd. 2006. Analysis of gene expression during development: lessons from the Apicomplexa. *Microbes Infect* 8 (6):1623-30.

Brennecke, J., C. D. Malone, A. A. Aravin, R. Sachidanandam, A. Stark, and G. J. Hannon. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 322 (5906):1387-92.

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

Bushati, N., and S. M. Cohen. 2007. microRNA functions. *Annu Rev Cell Dev Biol* 23:175-205.

Cai, Xuezhong, Curt H. Hagedorn, and Bryan R. Cullen. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10 (12):1957-1966.

Carlton, J. M., S. V. Angiuoli, B. B. Suh, T. W. Kooij, M. Pertea, J. C. Silva, M. D. Ermolaeva, J. E. Allen, J. D. Selengut, H. L. Koo, J. D. Peterson, M. Pop, D. S. Kosack, M. F. Shumway, S. L. Bidwell, S. J. Shallom, S. E. van Aken, S. B. Riedmuller, T. V. Feldblyum, J. K. Cho, J. Quackenbush, M. Sedegah, A. Shoaibi, L. M. Cummings, L. Florens, J. R. Yates, J. D. Raine, R. E. Sinden, M. A. Harris, D. A. Cunningham, P. R. Preiser, L. W. Bergman, A. B. Vaidya, L. H. van Lin, C. J. Janse, A. P. Waters, H. O. Smith, O. R. White, S. L. Salzberg, J. C. Venter, C. M. Fraser, S. L. Hoffman, M. J. Gardner, and D. J. Carucci. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* 419 (6906):512-9.

Carmell, M. A., Z. Xuan, M. Q. Zhang, and G. J. Hannon. 2002. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev* 16 (21):2733-42.

Cerutti, L., N. Mian, and A. Bateman. 2000. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem Sci* 25 (10):481-2.

Chapman, E. J., and J. C. Carrington. 2007. Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet* 8 (11):884-96.

Chendrimada, T. P., R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar. 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436 (7051):740-4.

Chendrimada, Thimmaiah P., Kenneth J. Finn, Xinjun Ji, David Baillat, Richard I. Gregory, Stephen A. Liebhaber, Amy E. Pasquinelli, and Ramin Shiekhattar. 2007. MicroRNA silencing through RISC recruitment of eIF6. *Nature* 447 (7146):823-828.

Chotivanich, K., R. Udomsangpetch, J. A. Simpson, P. Newton, S. Pukrittayakamee, S. Looareesuwan, and N. J. White. 2000. Parasite multiplication potential and the severity of Falciparum malaria. *J Infect Dis* 181 (3):1206-9.

Dannemann, B., J. A. McCutchan, D. Israelski, D. Antoniskis, C. Leport, B. Luft, J. Nussbaum, N. Clumeck, P. Morlat, J. Chiu, J.L. Vilde, M. Orellana, D. Feigal, A. Bartok, P. Heseltine, J. Leedom, and J. Remington. 1992. Treatment of Toxoplasmic Encephalitis in Patients with AIDS. *Annals of Internal Medicine* 116 (1):33-43.

Denli, A. M., B. B. Tops, R. H. Plasterk, R. F. Ketting, and G. J. Hannon. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature* 432 (7014):231-5.

Derry, M. C., A. Yanagiya, Y. Martineau, and N. Sonenberg. 2006. Regulation of poly(A)-binding protein through PABP-interacting proteins. *Cold Spring Harb Symp Quant Biol* 71:537-43.

Ding, L., A. Spencer, K. Morita, and M. Han. 2005. The developmental timing regulator AIN-1 interacts with miRISCs and may target the argonaute protein ALG-1 to cytoplasmic P bodies in C. elegans. *Mol Cell* 19 (4):437-47.

Doench, J. G., and P. A. Sharp. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev* 18 (5):504-11.

Dondorp, A. M., V. Desakorn, W. Pongtavornpinyo, D. Sahassananda, K. Silamut, K. Chotivanich, P. N. Newton, P. Pitisuttithum, A. M. Smithyman, N. J. White, and N. P. Day. 2005. Estimation of the total parasite biomass in acute falciparum malaria from plasma PfHRP2. *PLoS Med* 2 (8):e204.

Dubey, J. P. 2004. Toxoplasmosis – a waterborne zoonosis. *Vet Parasitol.* 126 (1-2):57-72.

Dubey, J. P. 2008. The history of Toxoplasma gondii--the first 100 years. *J Eukaryot Microbiol* 55 (6):467-75.

Elbashir, S. M., J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl. 2001. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411 (6836):494-8.

Elbashir, Sayda M., Winfried Lendeckel, and Thomas Tuschl. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development* 15 (2):188-200.

Eulalio, A., I. Behm-Ansmant, and E. Izaurralde. 2007. P bodies: at the crossroads of post-transcriptional pathways. *Nat Rev Mol Cell Biol* 8 (1):9-22.

Filipowicz, W., S. N. Bhattacharyya, and N. Sonenberg. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 9 (2):102-14.

Filippov, V., V. Solovyev, M. Filippova, and S. S. Gill. 2000. A novel type of RNase III family proteins in eukaryotes. *Gene* 245 (1):213-21.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 391 (6669):806-11.

Forstemann, K., Y. Tomari, T. Du, V. V. Vagin, A. M. Denli, D. P. Bratu, C. Klattenhoff, W. E. Theurkauf, and P. D. Zamore. 2005. Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol* 3 (7):e236.

Fortin, K. R., R. H. Nicholson, and A. W. Nicholson. 2002. Mouse ribonuclease III. cDNA structure, expression analysis, and chromosomal location. *BMC Genomics* 3 (1):26.

Gan, Jianhua, Joseph E. Tropea, Brian P. Austin, Donald L. Court, David S. Waugh, and Xinhua Ji. 2006. Structural Insight into the Mechanism of Double-Stranded RNA Processing by Ribonuclease III. 124 (2):355-366.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419 (6906):498-511.

Gardner, P. P., and R. Giegerich. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 5:140.

Gatignol, A., A. Buckler-White, B. Berkhout, and K. T. Jeang. 1991. Characterization of a human TAR RNA-binding protein that activates the HIV-1 LTR. *Science* 251 (5001):1597-600.

Giegerich, R., B. Voss, and M. Rehmsmeier. 2004. Abstract shapes of RNA. *Nucleic Acids Res* 32 (16):4843-51.

Giraldez, A. J., Y. Mishima, J. Rihel, R. J. Grocock, S. Van Dongen, K. Inoue, A. J. Enright, and A. F. Schier. 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* 312 (5770):75-9.

Goad, W. B., and M. I. Kanehisa. 1982. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Res* 10 (1):247-63.

Goldberg, A., and D. Robson. 1983. *Smalltalk-80: the language and its implementation*: Addison-Wesley Longman Publishing Co., Inc.

Gregory, R. I., T. P. Chendrimada, N. Cooch, and R. Shiekhattar. 2005. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* 123 (4):631-40.

Gregory, R. I., K. P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar. 2004. The Microprocessor complex mediates the genesis of microRNAs. *Nature* 432 (7014):235-40.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* 32 (Database issue):D109-11.

Griffiths-Jones, S., R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34 (Database issue):D140-4.

Griffiths-Jones, S., H. K. Saini, S. van Dongen, and A. J. Enright. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* 36 (Database issue):D154-8.

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

Grishok, A., J. L. Sinskey, and P. A. Sharp. 2005. Transcriptional silencing of a transgene by RNAi in the soma of C. elegans. *Genes Dev* 19 (6):683-96.

Gubbels, M. J., M. White, and T. Szatanek. 2008. The cell cycle and Toxoplasma gondii cell division: tightly knit or loosely stitched? *Int J Parasitol* 38 (12):1343-58.

Hammond, Scott M., Sabrina Boettcher, Amy A. Caudy, Ryuji Kobayashi, and Gregory J. Hannon. 2001. Argonaute2, a Link Between Genetic and Biochemical Analyses of RNAi. *Science* 293 (5532):1146-1150.

Han, J., Y. Lee, K. H. Yeom, Y. K. Kim, H. Jin, and V. N. Kim. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* 18 (24):3016-27.

Han, J., Y. Lee, K. H. Yeom, J. W. Nam, I. Heo, J. K. Rhee, S. Y. Sohn, Y. Cho, B. T. Zhang, and V. N. Kim. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* 125 (5):887-901.

Hill, D., and J. P. Dubey. 2002. Toxoplasma gondii: transmission, diagnosis and prevention. *Clin Microbiol Infect* 8 (10):634-40.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* 125 (2):167-188.

Houbaviy, H. B., M. F. Murray, and P. A. Sharp. 2003. Embryonic stem cell-specific MicroRNAs. *Dev Cell* 5 (2):351-8.

Humphreys, D. T., B. J. Westman, D. I. Martin, and T. Preiss. 2005. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proc Natl Acad Sci U S A* 102 (47):16961-6.

Hutvagner, G., J. McLachlan, A. E. Pasquinelli, E. Balint, T. Tuschl, and P. D. Zamore. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* 293 (5531):834-8.

Hutvagner, G., and P. D. Zamore. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* 297 (5589):2056-60.

IUPAC-IUB commission on biochemical nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. 1971. *J Mol Biol* 55 (3):299-310.

Jackson, R. J. 2005. Alternative mechanisms of initiating translation of mammalian mRNAs. *Biochem Soc Trans* 33 (Pt 6):1231-41.

Jaskiewicz, L., and W. Filipowicz. 2008. Role of Dicer in posttranscriptional RNA silencing. *Curr Top Microbiol Immunol* 320:77-97.

Ji, X. 2008. The mechanism of RNase III action: how dicer dices. *Curr Top Microbiol Immunol* 320:99-116.

Jones-Rhoades, M. W., D. P. Bartel, and B. Bartel. 2006. MicroRNAs and Their Regulatory Roles in Plants. *Annual Review of Plant Biology* 57 (1):19-53.

Kapp, L. D., and J. R. Lorsch. 2004. The molecular mechanics of eukaryotic translation. *Annu Rev Biochem* 73:657-704.

Ketting, R. F., S. E. J. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. A. Plasterk. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & Development* 15 (20):2654-2659.

Khvorova, A., A. Reynolds, and S. D. Jayasena. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115 (2):209-16.

Kim, J., A. Krichevsky, Y. Grad, G. D. Hayes, K. S. Kosik, G. M. Church, and G. Ruvkun. 2004. Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proc Natl Acad Sci U S A* 101 (1):360-5.

Kim, K., and L. M. Weiss. 2008. Toxoplasma: the next 100years. *Microbes Infect* 10 (9):978-84.

Kim, V. N. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* 6 (5):376-85.

Kim, V. N., and J. W. Nam. 2006. Genomics of microRNA. *Trends Genet* 22 (3):165-73.

Kiriakidou, M., G. S. Tan, S. Lamprinaki, M. De Planell-Saguer, P. T. Nelson, and Z. Mourelatos. 2007. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* 129 (6):1141-51.

Kissinger, J. C., B. Gajria, L. Li, I. T. Paulsen, and D. S. Roos. 2003. ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res* 31 (1):234-6.

Klattenhoff, C., and W. Theurkauf. 2008. Biogenesis and germline functions of piRNAs. *Development* 135 (1):3-9.

Knight, S. W., and B. L. Bass. 2001. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* 293 (5538):2269-71.

Kooij, T. W., C. J. Janse, and A. P. Waters. 2006. Plasmodium post-genomics: better the bug you know? *Nat Rev Microbiol* 4 (5):344-57.

Kruger, J., and M. Rehmsmeier. 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 34 (Web Server issue):W451-4.

Kurihara, Y., and Y. Watanabe. 2004. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A* 101 (34):12753-8.

Kurreck, J. 2009. RNA interference: from basic research to therapeutic applications. *Angew Chem Int Ed Engl* 48 (8):1378-98.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294 (5543):853-8.

Lagos-Quintana, M., R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* 12 (9):735-9.

Landthaler, M., A. Yalcin, and T. Tuschl. 2004. The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol* 14 (23):2162-7.

Lau, N. C., L. P. Lim, E. G. Weinstein, and D. P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 294 (5543):858-62.

Lee, R. C., C. M. Hammell, and V. Ambros. 2006. Interacting endogenous and exogenous RNAi pathways in Caenorhabditis elegans. *RNA* 12 (4):589-97.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V. N. Kim. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425 (6956):415-9.

Lee, Y., I. Hur, S. Y. Park, Y. K. Kim, M. R. Suh, and V. N. Kim. 2006. The role of PACT in the RNA silencing pathway. *EMBO J* 25 (3):522-32.

Lee, Y., K. Jeon, J. T. Lee, S. Kim, and V. N. Kim. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* 21 (17):4663-70.

Lee, Y., M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23 (20):4051-60.

Lee, Y. S., K. Nakahara, J. W. Pham, K. Kim, Z. He, E. J. Sontheimer, and R. W. Carthew. 2004. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* 117 (1):69-81.

Leuschner, P. J., S. L. Ameres, S. Kueng, and J. Martinez. 2006. Cleavage of the siRNA passenger strand during RISC assembly in human cells. *EMBO Rep* 7 (3):314-20.

Lewis, B. P., C. B. Burge, and D. P. Bartel. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120 (1):15-20.

Lim, L. P., N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel. 2003. The microRNAs of Caenorhabditis elegans. *Genes Dev* 17 (8):991-1008.

Lippman, Z., and R. Martienssen. 2004. The role of RNA interference in heterochromatic silencing. *Nature* 431 (7006):364-70.

Liu, J., F. V. Rivas, J. Wohlschlegel, J. R. Yates, 3rd, R. Parker, and G. J. Hannon. 2005. A role for the P-body component GW182 in microRNA function. *Nat Cell Biol* 7 (12):1261-6.

Liu, Qinghua, Tim A. Rand, Savitha Kalidas, Fenghe Du, Hyun-Eui Kim, Dean P. Smith, and Xiaodong Wang. 2003. R2D2, a Bridge Between the Initiation and Effector Steps of the Drosophila RNAi Pathway. *Science* 301 (5641):1921-1925.

Lu, C., D. H. Jeong, K. Kulkarni, M. Pillay, K. Nobuta, R. German, S. R. Thatcher, C. Maher, L. Zhang, D. Ware, B. Liu, X. Cao, B. C. Meyers, and P. J. Green. 2008. Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). *Proc Natl Acad Sci U S A* 105 (12):4951-6.

Lund, E., S. Guttinger, A. Calado, J. E. Dahlberg, and U. Kutay. 2004. Nuclear export of microRNA precursors. *Science* 303 (5654):95-8.

MacRae, I. J., K. Zhou, and J. A. Doudna. 2007. Structural determinants of RNA recognition and cleavage by Dicer. *Nat Struct Mol Biol* 14 (10):934-40.

Macrae, I. J., K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna. 2006. Structural basis for double-stranded RNA processing by Dicer. *Science* 311 (5758):195-8.

Malhotra, P., P. V. Dasaradhi, A. Kumar, A. Mohmmed, N. Agrawal, R. K. Bhatnagar, and V. S. Chauhan. 2002. Double-stranded RNA-mediated gene silencing of cysteine proteases (falcipain-1 and -2) of Plasmodium falciparum. *Mol Microbiol* 45 (5):1245-54.

Maniataki, E., and Z. Mourelatos. 2005. A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes Dev* 19 (24):2979-90.

Maroney, Patricia A., Yang Yu, Jesse Fisher, and Timothy W. Nilsen. 2006. Evidence that microRNAs are associated with translating messenger RNAs in human cells. *Nat Struct Mol Biol* 13 (12):1102-1107.

Martinez, J., A. Patkaniowska, H. Urlaub, R. Luhrmann, and T. Tuschl. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* 110 (5):563-74.

Matranga, Christian, Yukihide Tomari, Chanseok Shin, David P. Bartel, and Phillip D. Zamore. 2005. Passenger-Strand Cleavage Facilitates Assembly of siRNA into Ago2-Containing RNAi Enzyme Complexes. 123 (4):607-620.

McRobert, L., and G. A. McConkey. 2002. RNA interference (RNAi) inhibits growth of Plasmodium falciparum. *Mol Biochem Parasitol* 119 (2):273-8.

Meins, F., Jr., A. Si-Ammour, and T. Blevins. 2005. RNA silencing systems and their relevance to plant development. *Annu Rev Cell Dev Biol* 21:297-318.

Meissner, M., M. S. Breinich, P. R. Gilson, and B. S. Crabb. 2007. Molecular genetic tools in Toxoplasma and Plasmodium: achievements and future needs. *Curr Opin Microbiol* 10 (4):349-56.

Meister, G., M. Landthaler, L. Peters, P. Y. Chen, H. Urlaub, R. Luhrmann, and T. Tuschl. 2005. Identification of novel argonaute-associated proteins. *Curr Biol* 15 (23):2149-55.

Meister, G., and T. Tuschl. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* 431 (7006):343-9.

Mello, C. C., and D. Conte, Jr. 2004. Revealing the world of RNA interference. *Nature* 431 (7006):338-42.

Merrick, W. C. 2004. Cap-dependent and cap-independent translation in eukaryotic systems. *Gene* 332:1-11.

Minsky, M. . 1974. A Framework for Representing Knowledge: Massachusetts Institute of Technology.

miRBase. 2009. miRBase: the microRNA database. http://www.mirbase.org/ (accessed December 6, 2009).

Mishima, Y., A. J. Giraldez, Y. Takeda, T. Fujiwara, H. Sakamoto, A. F. Schier, and K. Inoue. 2006. Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. *Curr Biol* 16 (21):2135-42.

Miyoshi, K., H. Tsukumo, T. Nagami, H. Siomi, and M. C. Siomi. 2005. Slicer function of Drosophila Argonautes and its involvement in RISC formation. *Genes Dev* 19 (23):2837-48.

Mohmmed, A., P. V. Dasaradhi, R. K. Bhatnagar, V. S. Chauhan, and P. Malhotra. 2003. In vivo gene silencing in Plasmodium berghei--a mouse malaria model. *Biochem Biophys Res Commun* 309 (3):506-11.

Mootz, D., D. M. Ho, and C. P. Hunter. 2004. The STAR/Maxi-KH domain protein GLD-1 mediates a developmental switch in the translational control of C. elegans PAL-1. *Development* 131 (14):3263-72.

Moss, E. G. 2001. RNA interference: it's a small RNA world. *Curr Biol* 11 (19):R772-5.

Mourelatos, Z., J. Dostie, S. Paushkin, A. Sharma, B. Charroux, L. Abel, J. Rappsilber, M. Mann, and G. Dreyfuss. 2002. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* 16 (6):720-8.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48 (3):443-53.

Nelson, P. T., A. G. Hatzigeorgiou, and Z. Mourelatos. 2004. miRNP:mRNA association in polyribosomes in a human neuronal cell line. *RNA* 10 (3):387-94.

Nicolle, C., and L. Manceaux. 1908. Sur une infection a corps de Leishman (ouorganisms voisons) du gondi. *C.R. Acad. Sci.* 148:763-766.

Nielsen, C. B., N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge. 2007. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* 13 (11):1894-910.

Nottrott, S., M. J. Simard, and J. D. Richter. 2006. Human let-7a miRNA blocks protein production on actively translating polyribosomes. *Nat Struct Mol Biol* 13 (12):1108-14.

Nowotny, M., and W. Yang. 2009. Structural and functional modules in RNA interference. *Curr Opin Struct Biol* 19 (3):286-93.

O'Donnell, K. A., and J. D. Boeke. 2007. Mighty Piwis defend the germline against genome intruders. *Cell* 129 (1):37-44.

Park, W., J. Li, R. Song, J. Messing, and X. Chen. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol* 12 (17):1484-95.

Parker, R., and U. Sheth. 2007. P bodies and the control of mRNA translation and degradation. *Mol Cell* 25 (5):635-46.

Parker, R., and H. Song. 2004. The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol* 11 (2):121-7.

Pasquinelli, A. E., B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Muller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408 (6808):86-9.

Peragine, A., M. Yoshikawa, G. Wu, H. L. Albrecht, and R. S. Poethig. 2004. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev* 18 (19):2368-79.

Petersen, C. P., M. E. Bordeleau, J. Pelletier, and P. A. Sharp. 2006. Short RNAs repress translation after initiation in mammalian cells. *Mol Cell* 21 (4):533-42.

Pham, J. W., J. L. Pellino, Y. S. Lee, R. W. Carthew, and E. J. Sontheimer. 2004. A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi in Drosophila. *Cell* 117 (1):83-94.

Pillai, R. S., S. N. Bhattacharyya, C. G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, and W. Filipowicz. 2005. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science* 309 (5740):1573-6.

Preall, J. B., and E. J. Sontheimer. 2005. RNAi: RISC gets loaded. *Cell* 123 (4):543-5.

Radke, J. R., B. Striepen, M. N. Guerini, M. E. Jerome, D. S. Roos, and M. W. White. 2001. Defining the cell cycle for the tachyzoite stage of Toxoplasma gondii. *Mol Biochem Parasitol* 115 (2):165-75.

Radke, Jay, Michael Behnke, Aaron Mackey, Josh Radke, David Roos, and Michael White. 2005. The transcriptome of Toxoplasma gondii. *BMC Biology* 3 (1):26.

Rana, T. M. 2007. Illuminating the silence: understanding the structure and function of small RNAs. *Nat Rev Mol Cell Biol* 8 (1):23-36.

Rand, T. A., S. Petersen, F. Du, and X. Wang. 2005. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell* 123 (4):621-9.

Reeder, J., and R. Giegerich. 2005. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* 21 (17):3516-23.

Rehmsmeier, M., P. Steffen, M. Hochsmann, and R. Giegerich. 2004. Fast and effective prediction of microRNA/target duplexes. *RNA* 10 (10):1507-17.

Reilly, H. B., H. Wang, J. A. Steuter, A. M. Marx, and M. T. Ferdig. 2007. Quantitative dissection of clone-specific growth rates in cultured malaria parasites. *Int J Parasitol* 37 (14):1599-607.

Ro, S., C. Park, D. Young, K. M. Sanders, and W. Yan. 2007. Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res* 35 (17):5944-53.

Rodriguez, A., S. Griffiths-Jones, J. L. Ashurst, and A. Bradley. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res* 14 (10A):1902-10.

Rose, S. D., D. H. Kim, M. Amarzguioui, J. D. Heidel, M. A. Collingwood, M. E. Davis, J. J. Rossi, and M. A. Behlke. 2005. Functional polarity is introduced by Dicer processing of short substrate RNAs. *Nucl. Acids Res.* 33 (13):4140-4156.

Ruby, J. Graham, Calvin H. Jan, and David P. Bartel. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* 448 (7149):83-86.

Ruegsegger, U., J. H. Leber, and P. Walter. 2001. Block of HAC1 mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. *Cell* 107 (1):103-14.

Sabin, Albert B., and Peter K. Olitsky. 1937. *Toxoplasma* and obligate intracellular parasitism. *Science* 85 (2205):336-338.

Saito, K., A. Ishizuka, H. Siomi, and M. C. Siomi. 2005. Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. *PLoS Biol* 3 (7):e235.

Sankoff, David. 1985. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics* 45 (5):810-825.

Sanvito, F., S. Piatti, A. Villa, M. Bossi, G. Lucchini, P. C. Marchisio, and S. Biffo. 1999. The beta4 integrin interactor p27(BBP/eIF6) is an essential nuclear matrix protein involved in 60S ribosomal subunit assembly. *J Cell Biol* 144 (5):823-37.

Sasaki, T., A. Shiohama, S. Minoshima, and N. Shimizu. 2003. Identification of eight members of the Argonaute family in the human genome small star, filled. *Genomics* 82 (3):323-30.

Schmitter, D., J. Filkowski, A. Sewer, R. S. Pillai, E. J. Oakeley, M. Zavolan, P. Svoboda, and W. Filipowicz. 2006. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res* 34 (17):4801-15.

Schwarz, D. S., G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115 (2):199-208.

Shi, H., A. Djikeng, T. Mark, E. Wirtz, C. Tschudi, and E. Ullu. 2000. Genetic interference in Trypanosoma brucei by heritable and inducible double-stranded RNA. *RNA* 6 (7):1069-76.

Si, K., and U. Maitra. 1999. The Saccharomyces cerevisiae homologue of mammalian translation initiation factor 6 does not function as a translation initiation factor. *Mol Cell Biol* 19 (2):1416-26.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* 147 (1):195-7.

Steffen, P., B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22 (4):500-3.

Stefik, M. , and D. G. Bobrow. 1986. Object-oriented programming: Themes and variations. *AI Mag.* 6 (4):40-62.

Striepen, B., C. N. Jordan, S. Reiff, and G. G. van Dooren. 2007. Building the perfect parasite: cell division in apicomplexa. *PLoS Pathog* 3 (6):e78.

Sun Microsystems, Inc. 2009. The Java Tutorials. http://java.sun.com/docs/books/tutorial (accessed November 29, 2009).

Tahbaz, N., F. A. Kolb, H. Zhang, K. Jaronczyk, W. Filipowicz, and T. C. Hobman. 2004. Characterization of the interactions between mammalian PAZ PIWI domain proteins and Dicer. *EMBO Rep* 5 (2):189-94.

Taylor, S., A. Barragan, C. Su, B. Fux, S. J. Fentress, K. Tang, W. L. Beatty, H. E. Hajj, M. Jerome, M. S. Behnke, M. White, J. C. Wootton, and L. D. Sibley. 2006. A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen Toxoplasma gondii. *Science* 314 (5806):1776-80.

Till, S., E. Lejeune, R. Thermann, M. Bortfeld, M. Hothorn, D. Enderle, C. Heinrich, M. W. Hentze, and A. G. Ladurner. 2007. A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nat Struct Mol Biol* 14 (10):897-903.

Timms, R., N. Colegrave, B. H. Chan, and A. F. Read. 2001. The effect of parasite dose on disease severity in the rodent malaria Plasmodium chabaudi. *Parasitology* 123 (1):1-11.

Tolia, N. H., and L. Joshua-Tor. 2007. Slicer and the argonautes. *Nat Chem Biol* 3 (1):36-43.

Tomari, Y., T. Du, B. Haley, D. S. Schwarz, R. Bennett, H. A. Cook, B. S. Koppetsch, W. E. Theurkauf, and P. D. Zamore. 2004. RISC assembly defects in the Drosophila RNAi mutant armitage. *Cell* 116 (6):831-41.

Tomari, Y., T. Du, and P. D. Zamore. 2007. Sorting of Drosophila small silencing RNAs. *Cell* 130 (2):299-308.

Tomari, Y., C. Matranga, B. Haley, N. Martinez, and P. D. Zamore. 2004. A protein sensor for siRNA asymmetry. *Science* 306 (5700):1377-80.

Tomari, Y., and P. D. Zamore. 2005. Perspective: machines for RNAi. *Genes Dev* 19 (5):517-29.

ToxoDB. 2009. *Toxoplasma gondii* Genome Resource. http://toxodb.org/toxo/ (accessed November 30, 2009).

Ullu, E., C. Tschudi, and T. Chakraborty. 2004. RNA interference in protozoan parasites. *Cell Microbiol* 6 (6):509-19.

Vasudevan, S., and J. A. Steitz. 2007. AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell* 128 (6):1105-18.

Vaucheret, H. 2006. Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev* 20 (7):759-71.

Vazquez, F., H. Vaucheret, R. Rajagopalan, C. Lepers, V. Gasciolli, A. C. Mallory, J. L. Hilbert, D. P. Bartel, and P. Crete. 2004. Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol Cell* 16 (1):69-79.

Verdel, A., S. Jia, S. Gerber, T. Sugiyama, S. Gygi, S. I. Grewal, and D. Moazed. 2004. RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* 303 (5658):672-6.

Voss, B., R. Giegerich, and M. Rehmsmeier. 2006. Complete probabilistic analysis of RNA shapes. *BMC Biol* 4:5.

Wang, B., T. M. Love, M. E. Call, J. G. Doench, and C. D. Novina. 2006. Recapitulation of short RNA-directed translational gene silencing in vitro. *Mol Cell* 22 (4):553-60.

Wang, Zefeng, James C. Morris, Mark E. Drew, and Paul T. Englund. 2000. Inhibition of Trypanosoma brucei Gene Expression by RNA Interference Using an Integratable Vector with Opposing T7 Promoters. *Journal of Biological Chemistry* 275 (51):40174-40179.

Weinreb, D. . 1981. *Lisp machine manual*: Massachusetts Institute of Technology.

Wells, S. E., P. E. Hillner, R. D. Vale, and A. B. Sachs. 1998. Circularization of mRNA by eukaryotic translation initiation factors. *Mol Cell* 2 (1):135-40.

Wu, L., and J. G. Belasco. 2005. Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. *Mol Cell Biol* 25 (21):9198-208.

Wu, L., J. Fan, and J. G. Belasco. 2006. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A* 103 (11):4034-9.

Xie, Z., E. Allen, A. Wilken, and J. C. Carrington. 2005. DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 102 (36):12984-9.

Xie, Z., L. K. Johansen, A. M. Gustafson, K. D. Kasschau, A. D. Lellis, D. Zilberman, S. E. Jacobsen, and J. C. Carrington. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* 2 (5):E104.

Yekta, S., I. H. Shih, and D. P. Bartel. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304 (5670):594-6.

Yi, R., Y. Qin, I. G. Macara, and B. R. Cullen. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* 17 (24):3011-6.

Zamore, P. D., and B. Haley. 2005. Ribo-gnome: the big world of small RNAs. *Science* 309 (5740):1519-24.

Zamore, Phillip D. 2004. Plant RNAi: How aViral Silencing Suppressor Inactivates siRNA. 14 (5):R198-R200.

Zeng, Y., and B. R. Cullen. 2005. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J Biol Chem* 280 (30):27595-603.

Zeng, Y., R. Yi, and B. R. Cullen. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J* 24 (1):138-48.

Zhang, B., X. Pan, C. H. Cannon, G. P. Cobb, and T. A. Anderson. 2006. Conservation and divergence of plant microRNA genes. *Plant J* 46 (2):243-59.

Zhang, H., F. A. Kolb, V. Brondani, E. Billy, and W. Filipowicz. 2002. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J* 21 (21):5875-5885.

Zhang, H., F. A. Kolb, L. Jaskiewicz, E. Westhof, and W. Filipowicz. 2004. Single processing center models for human Dicer and bacterial RNase III. *Cell* 118 (1):57-68.

APPENDIX-A

VALUES THAT ARE CALCULATED FOR HAIRPINS FROM MIRBASE

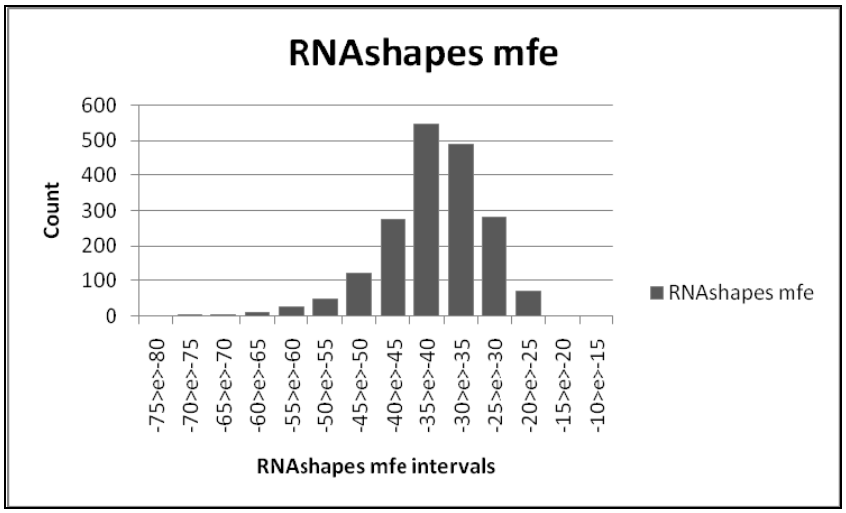

Figure A.1 Minimum free energy values of folded hairpins. Hairpins are folded by RNAshapes program as in our system. Horizontal Axis denotes intervals of mfe (kcal/mol) while vertical axis denotes number of elements in each interval.

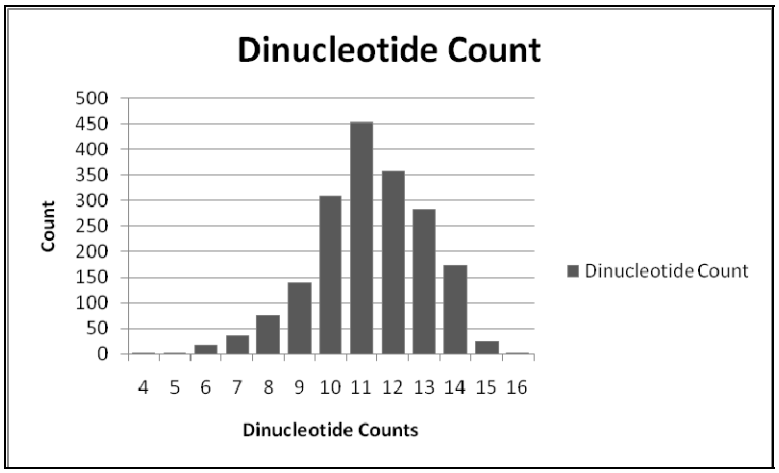

Figure A.2 Dinucleotide counts in mature miRNAs obtained from hairpins. Horizontal axis denotes number of different dinucleotides in miRNAs while vertical axis denotes number of miRNAs for each dinucleotide number.

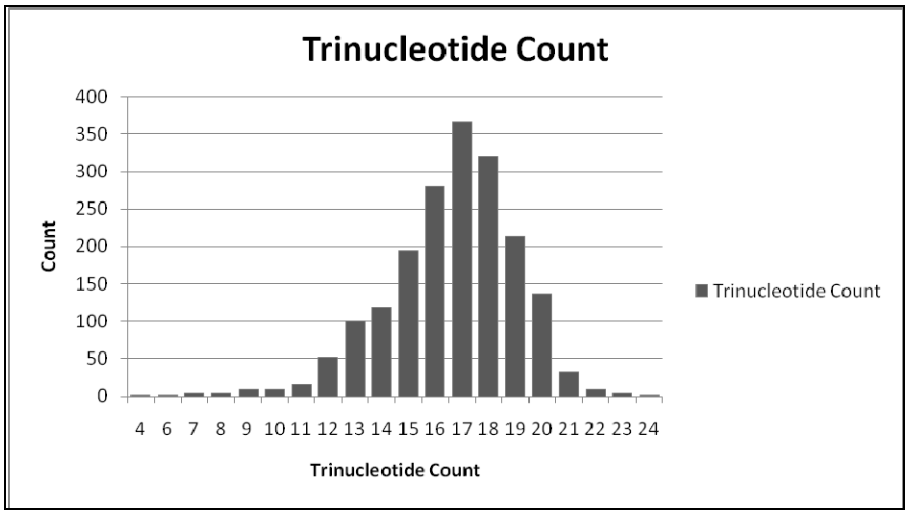

Figure A.3 Trinucleotide counts in mature miRNAs obtained from hairpins. Horizontal axis denotes number of different trinucleotides in miRNAs while vertical axis denotes number of miRNAs for each dinucleotide number.

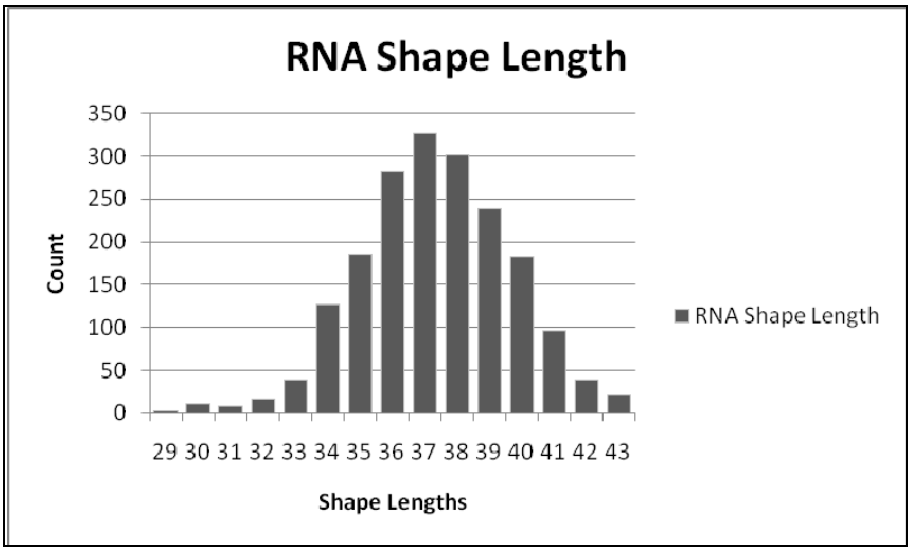

Figure A.4 Shape lengths of hairpins after folding by RNAshapes. Horizontal axis denotes length of shape representations while vertical axis denotes number of hairpins with corresponding length.

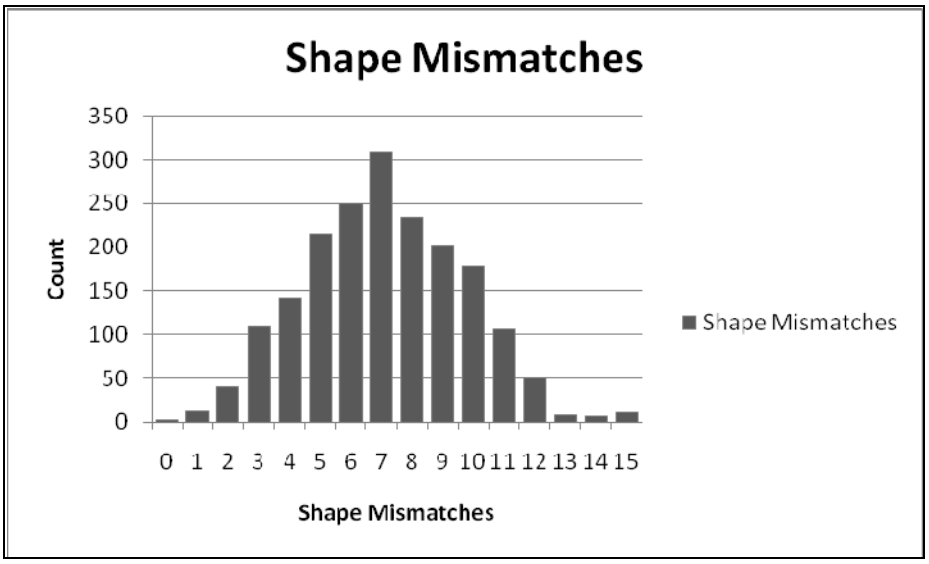

Figure A.5 Mismatches in shapes after folding by RNAshapes. Horizontal axis denotes the number of mismatches in shape representations while vertical axis denotes number of hairpins with corresponding number of mismatches.

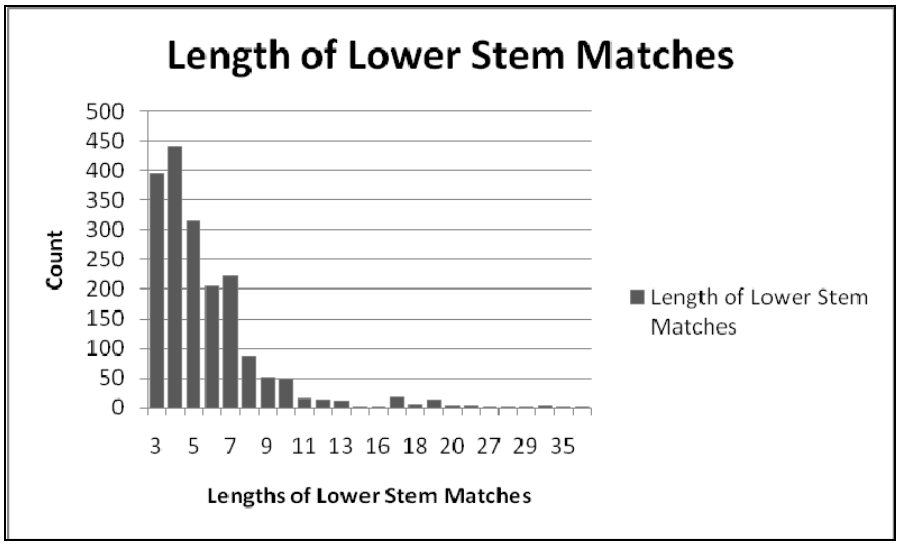

Figure A.6 Length of lower stem matches in folded hairpins. Horizontal axis denotes the number of matches in the lower stem. Lower stem is the region from SD junction to closes mismatch. Vertical axis denotes the number of hairpins with corresponding length of lower stem matches.

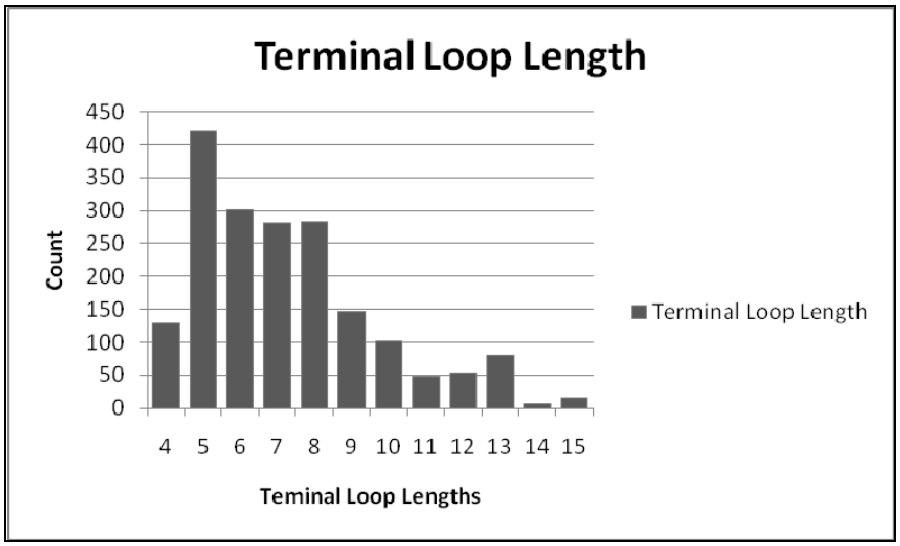

Figure A.7 Length of the terminal loops in folded hairpins. Horizontal axis denotes the length of terminal loops while vertical axis denotes number of hairpins with corresponding terminal loop length.

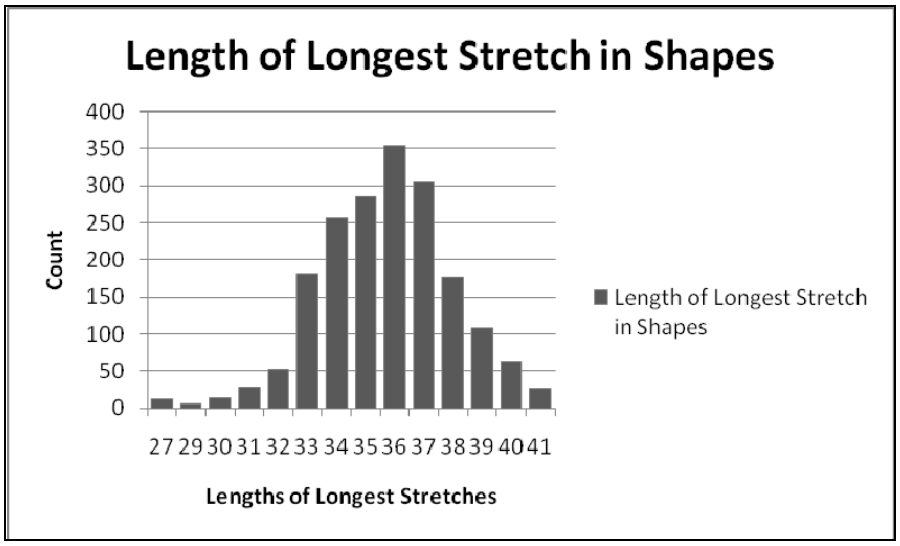

Figure A.8 Length of longest stretches in hairpins. Longest stretch is the stretch from SD junction to the terminal loop. Horizontal axis denotes the length of longest stretch while vertical axis denotes the number of hairpins with corresponding stretch length.

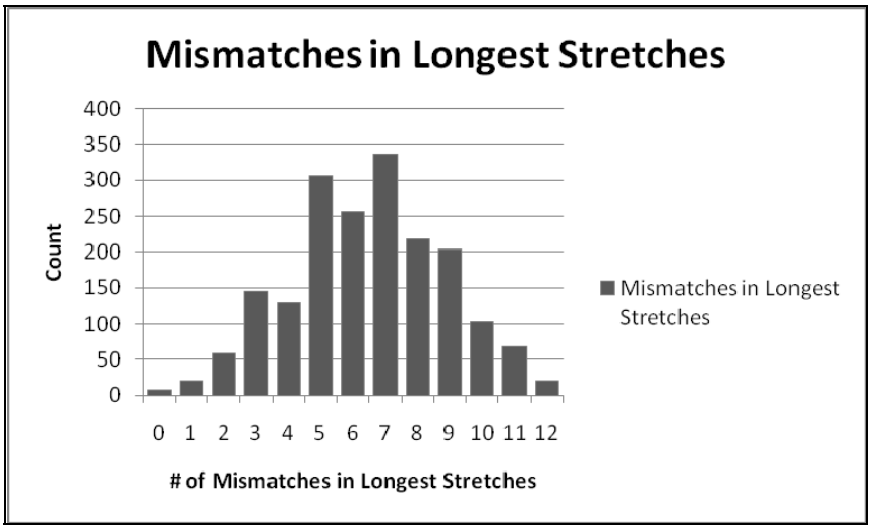

Figure A.9 Number of mismatches in longest stretches. Horizontal axis denotes the number of mismatched nucleotides in longest stretch while vertical axis denotes the number of hairpins with corresponding longest stretch mismatch number.

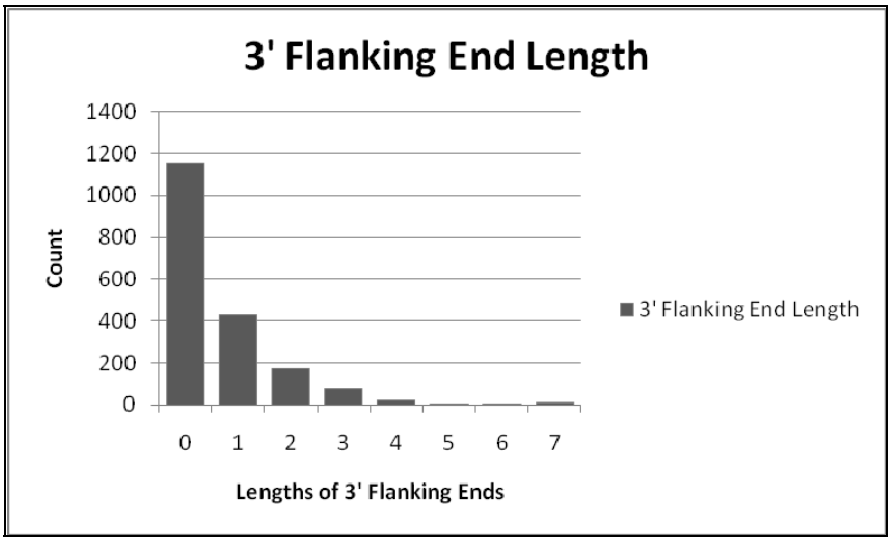

Figure A.10 Length of the flanking end on 3’ end of hairpin. Horizontal axis denotes the length of 3’ flanking end while vertical axis denotes number of hairpins with corresponding 3’ flanking end length.

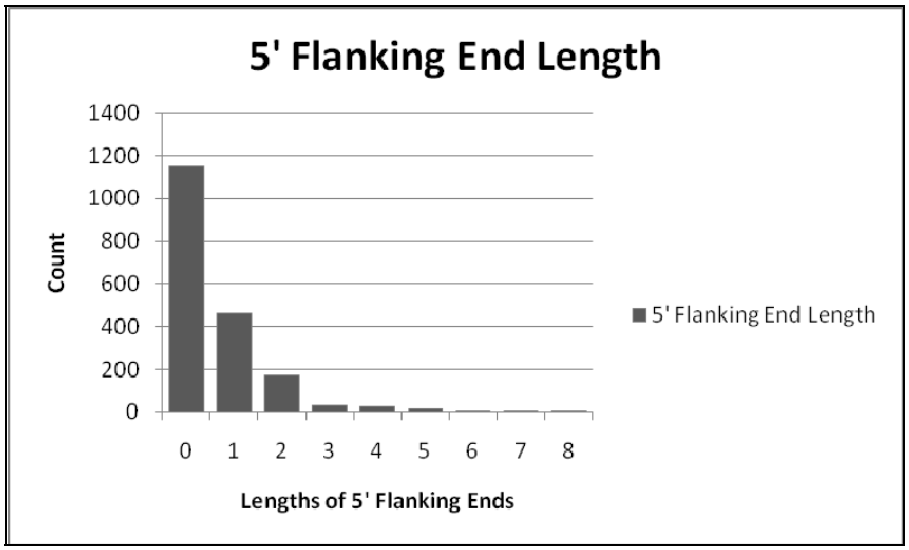

Figure A.11 Length of the flanking end on 5’ end of hairpin. Horizontal axis denotes the length of 5’ flanking end while vertical axis denotes number of hairpins with corresponding 5’ flanking end length.