COMPUTATIONAL PREDICTION OF CHICKEN PRE MICRO RNA

by

BRAM SEBASTIAN

(Under the Direction of SAMUEL E. AGGREY)

ABSTRACT

MicroRNAs (miRNAs) are a small single strand non-coding RNA with ~22 nucleotides (nts) which can regulate gene expression. MiRNAs are generated from a ~60-70 nts long hairpin shapped pre-miRNA which is a product of the cleavage of primary miRNA. The slow pace of the identification of novel miRNA by laboratory experiment has raised the importance of computational method. Three major programs ProMir, ERPIN, and MiR-abela were tested for known chicken pre-miRNA. ProMir, ERPIN and MiR-abela detected 53%, 93% and 57% respectively where ERPIN only has 25% of the chicken miRNA classes available. Thus, novel computational approach miR-Explore is introduced which was demonstrated to have 89% sensitivity in identifying known chicken pre-miRNA.

INDEX WORDS: computational prediction, microRNA, chicken, miR-Explore

COMPUTATIONAL PREDICTION OF CHICKEN PRE MICRO RNA

by

BRAM SEBASTIAN

B.S. Iowa State University, 2004

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2007

COMPUTATIONAL PREDICTION OF CHICKEN PRE MICRO RNA

by

BRAM SEBASTIAN

Major Professor:	Samuel E. Aggrey
Committee:	Liming Cai Paul Schliekelman

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007

DEDICATION

I would like to dedicate this thesis to my father, my mother and my family for their supports during my study. Also to my late grandfather who encourage me to study in the United States but unfortunately did not have chance to witness my academic achievement.Thank you all.

ACKNOWLEDGEMENTS

I thank Dr. Samuel Aggrey, Dr. Liming Cai, and Dr. Paul Schliekelman for their time, help and support for me to be able to finish my master. I also thank Dr. Cai's graduate student Wu Yong for his help and suggestion to complete my master project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

MicroRNAs (miRNAs) are non-coding single stranded RNAs that regulate gene expression in the post transcriptional state by arresting the mRNA translation or by their cleavage (Bartel 2004). MiRNAs consist of ~22 nucleotides (nts) which are generated from endogenous transcripts that can form local hairpin structures (Kim 2005). MiRNA genes are transcribed by RNA polymerase II. The local hairpin structures in the primary transcripts are called pri-miRNAs which are first processed by Drosha (nuclear RNAse type III enzyme). After being processed by Drosha, the pri-miRNAs will release the hairpin shaped intermediates which are called pre-miRNA which have ~60-70 nts. Pre-miRNAs consist of ~22nt double stranded stem and a ~10nt loop. Pre-miRNAs are then transported from the nucleus to the cytoplasm by the nuclear export factor Exportin-5 and Ran-GTP cofactor (Bartel 2004) where they are cleaved by Dicer (RNAse III enzyme) to yield a ~22nt RNA duplex. One strand of this RNA duplex will be selected as mature miRNA which will be associated with the RNA-induced silencing complex (RISC) to target single stranded mRNA for degradation and the other will be destroyed.

The effort to identify miRNA started after the discovery of lin-4 and let-7 genes by analyzing the developmental timing of *Caenorhabditis elegans* (*C. elegans*) (Lee, Feinbaum et al. 1993; Reinhart, Slack et al. 2000). Since then, hundreds of miRNAs has been discovered in plants, animals and viruses, and deposited in an online database called

miRBase (http://microrna.sanger.ac.uk/sequences/) (Griffiths-Jones, Grocock et al. 2006). The identification of miRNA using wet lab experimental procedures such as cDNA cloning and small RNA cloning have been accomplished (Wang, Zhou et al. 2004; Samols, Hu et al. 2005). However, such procedures do not detect miRNAs expressed at low levels. Thus, computational approaches are developed to overcome at least, partially these problems.

The basic computational approach for this problem has been to train known sequences to find the unknown ones. The fact that the mechanism behind miRNA and their action are not completely known has made the computational task related to miRNA a very challenging problem. The computational miRNA prediction is often based on the miRNA biological features. Some of those features are: (1)The miRNA genes are small non-coding genes less than 150 bp that cloned several times with clone length between 21-23nt (Bartel 2004; Pfeffer, Zavolan et al. 2004), (2) miRNAs are normally conserved in the genomes of related species but some miRNAs may be conserved in all animals (Bartel 2004; Pfeffer, Zavolan et al. 2004), (3) miRNAs are often occur in cluster usually separated by some kilobases (Seitz, Youngson et al. 2003; Sachidanandam 2005; Sewer, Paul et al. 2005), (4) miRNA has to be the complement of the 3' UTRs of the target mRNA (Bartel 2004), and (5) miRNAs can exist in intergenic regions, introns of protein coding regions, or exons and introns of non-coding genes (Rodriguez, Griffiths-Jones et al. 2004).

Several computational methods to predict miRNA has been developed. The major programs are PalGrade (Bentwich, Avniel et al. 2005), MiRScan and Easy RNA profile identification (ERPIN) (Gautheret and Lambert 2001). Besides PalGrade, MiRScan and

ERPIN, several other computational miRNA predictors that have been developed (Zhang, Pan et al. 2006), but they can only detect miRNAs which are abundantly expressed or miRNAs that are homolog to the previous known miRNAs. To ameliorate this problem ProMir (Nam, Shin et al. 2005) was developed. ProMir was developed with a purpose of identifying human miRNA through a probabilistic co-learning model of sequence and structure. The method simultaneously considers the structure and sequence of the miRNA precursors. This method was improved by the development of ProMir II (Nam, Kim et al. 2006) which combine free energy data, G/C ratio, conservation score and entropy of candidate sequences for more accurate and controllable prediction.

Although many miRNA prediction algorithms have been developed for the discovery of several novel miRNAs, most of them were developed either for a specific species or based on some broad assumptions, and a result a great number of species specific miRNAs are still awaiting for discovery (Zhang, Pan et al. 2006). The objective of this thesis is to (1) test the efficacy of existing computational algorithm in detecting known chicken pre-miRNA, and (2) develop and alternative computational method to identify all known chicken miRNA and additional novel miRNA candidates in the chicken genome.

CHAPTER 2

LITTERATURE REVIEW

2.1 From DNA to RNA

Deoxyribonucleic acid (DNA) is a genetic material located in the nucleus of the cell which carries the genetic information of both prokaryotes and eukaryotes organisms. DNA is also the genetic material for some viruses but, some viruses have ribonucleic acid (RNA) as their genetic material. The full DNA sequence of an organism or full RNA sequence of some viruses is called a genome. DNA consists of two double helix strands. Each of the strands has some nucleotides, which contain the sugar deoxyribose, a phosphate group and a base (Watson and Crick 1953). The organization of these nucleotides is the cause of the double helix structure of the DNA (Figure 2.1) http://www.sc.chula.ac.th/courseware/2305262/context/text7/pic7/a_7-8.gif.

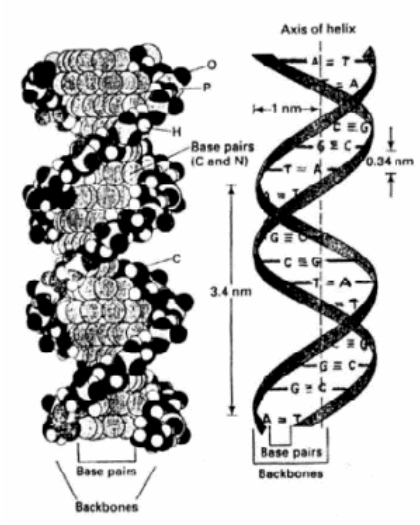

Figure 2.1 DNA Double Helix Structure

DNA has four bases which are Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). In RNA, Thymine is replaced by Uracil (U). The genetic information in the DNA strand determine by the sequence of this bases, and specific sequences of the nucleotides is known as genes. Genes encode proteins that are responsible for reproduction, disease, development and other significant functions. The processes of making the proteins consist of two main steps: transcription and translation. These two steps are then well known as the central dogma (DNA to RNA to proteins) which is introduced by Crick in 1958 and later restated in 1970 (Crick 1970). There are four major types of RNA. They are messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and small nuclear RNA (snRNA).

2.1.1 Transcription

Transcription process of eukaryotes and prokaryotes is dissimilar. In general, transcription is catalyzed by the enzyme RNA polymerase. RNA is transcribed in the 5' to 3' direction. The 3' to 5' DNA strand that is read to produce the RNA is called the template strand. In both eukaryotes and prokaryotes, transcription is started by initiation, followed by elongation and termination. In prokaryotes, the transcribed genes can be divided into three parts:

1. The promoter, the sequence upstream of the start of the DNA sequence that is transcribed to the RNA.
2. The transcribed DNA sequence which is well known as RNA coding sequence.
3. The terminator, the sequence signaling the transcription to stop.

There are two DNA sequences in the promoter region that are important to the transcription process. These sequences can be found in -35 and -10. The consensus sequence for the -35 position is 5'-TTGACA-3' and the consensus sequence for the -10 position is 5'-TATAAT-3' (previously known as Prinbow Box). To start the transcription process, a form of RNA polymerase called the holoenzyme binds to the promoter. Holoenzyme is bound with another polypeptide called sigma factor which is used to recognize the -35 and -10 region in the promoter. The holoenzyme first bind loosely to the -35 region and then bind more tightly to the -10 region followed by the untwisting of the DNA which form an open promoter complex. The rate of transcription varies due to the difference of each actual sequence of each promoter which can affect the binding efficiency of RNA polymerase (Russell 2006). This variation causes genes to have distinct level of expression from one to another. The untwisting continues to the region of the RNA coding sequence and form a transcription bubble. The synthesis of the RNA takes place in the transcription bubble. Only one strand of the DNA functions as the template. That is the 5' to 3' strand. Once about 10 RNA nucleotides have been formed, the sigma factor is released from the RNA polymerase to be used in other transcription initiation process. The core of the RNA polymerase moves on to continue the transcription process. It untwists the DNA double helix ahead of it and twists back the DNA behind it. This process continues until it reaches the termination sequence. In prokaryotes, there is only one RNA polymerase to synthesize mRNA, tRNA and rRNA (Borukhov and Nudler 2003).

Transcription is more complicated in the eukaryotes. In eukaryotes, there are three different RNA polymerase. RNA polymerase I synthesize rRNA and can be found in the

ribosome. RNA polymerase II is in the nucleoplasm of the nucleus and used to synthesize mRNA and some snRNA. RNA polymerase III synthesizes rRNA which are not synthesized by RNA polymerase I, tRNA, and the snRNA that are not synthesized by the RNA polymerase II. RNA polymerase III can be found only in the nucleoplasm. Unlike in prokaryotes, the detail of the RNA polymerase in eukaryotes is less known. In the eukaryotes, promoter can be divided into two regions, *the core promoter* and the *promoter proximal elements*. The core promoter consists of short sequence element called initiator and the TATA box, located at position -30. The TATA box has the consensus sequence TATAAAA. These two elements decide where all of the elements necessary for transcription assembled and also the starting position of the transcription.

Promoter proximal elements located in the regions of 50 to 200 nucleotides from the starting position of the transcription. Some examples are CAAT (cat box) which is located at -75 and the GGGCGG (GC box) centered at -90. These elements determine the efficiency of the promoter. Besides promoter, eukaryotes also have another sequence called enhancers to maximize the transcription of a gene. The location of enhancers is usually upstream or downstream from the gene and separated by thousands of basepairs (Harvey Lodish 2000). Thus, enhancers function from a distance.

The transcription of protein coding gene is initiated by the RNA polymerase II which works together with other proteins called general transcription factors (GTFs) (Harvey Lodish 2000). The building of the RNA polymerase II and the GTFs takes place in the core promoter. In eukaryotes, all three RNA polymerase required to work together with the GTFs to be able to function properly. The complex formed by the RNA polymerase II and some GTFs unwinds the promoter to start the transcription.

The product of the transcription of protein coding gene is mRNA. The molecule of the mRNA consist of 5' untranslated region (UTR), the coding sequence, and the 3'UTR. The general structure of mRNA for both prokaryotes and eukaryotes can be seen in Figure 2.2. Prokaryotes mRNA can contain the information of more than one gene, and eukaryotes mRNA can only contain the information of one gene. In prokaryotes, the RNA transcript function directly as an mRNA, whereas in the eukaryotes, the RNA transcript is called pre-mRNA and must be modified in the nucleus to produce mature mRNA.

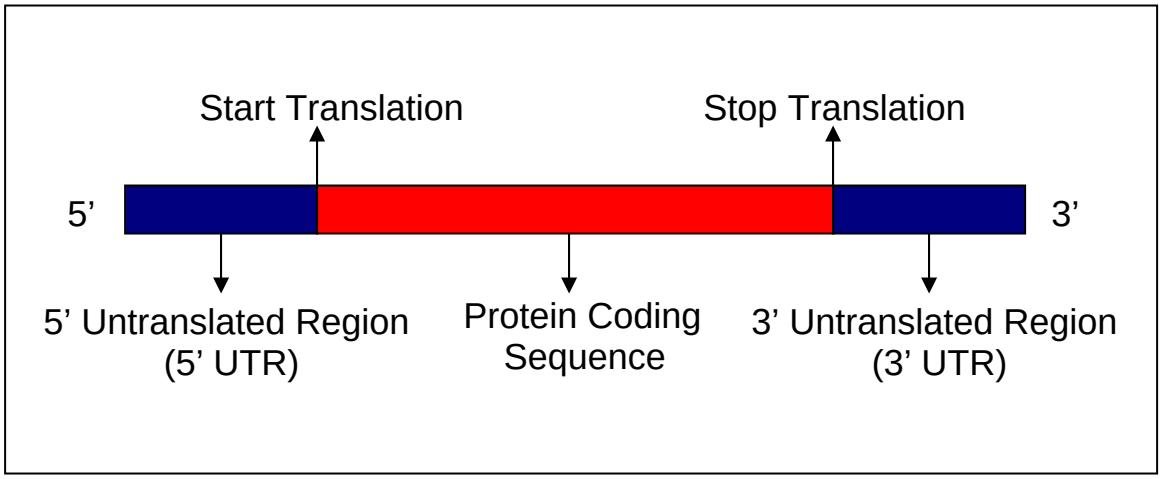

Figure 2.2 General Structure of mRNA

2.1.1.1 The Process of Producing Mature mRNA

In eukaryotes, the pre-mRNAs are modified in both 5' and 3' ends. In the protein coding sequence region of pre-mRNA, there are non amino acid coding sequence called introns and the sequences appears in the mature mRNA called exons (Gilbert 1978). The discovery of introns led to the Noble Price for Physiology in Medicine in 1993 for Phillip

Allen Sharp and Richard J. Roberts. The general structure of pre-mRNA can bee seen in Figure 2.3.

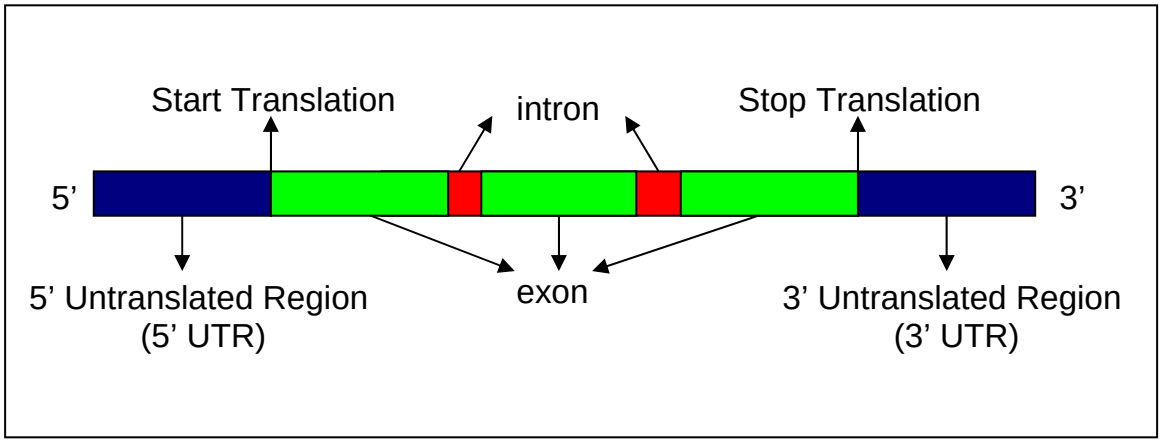

Figure 2.3 General structure of pre-mRNA

The processing of the pre-mRNA starts by the addition of guanine nucleotide by the capping enzyme (Lewis and Izaurralde 1997). This process is known as 5' capping. The 5' cap remains in place during the processing and protecting the mRNA from degradation. The 5' cap is also important for the mature mRNA to be able to bind to the ribosome to start translation. The next process is the modification of the 3' end by the addition of approximately 50 to 250 adenine nucleotides which is well known as poly A tail. The poly A tail is important for exporting the mRNA from the nucleus to the cytoplasm efficiently. Once in the cytoplasm, the poly A tail protect the mRNA's 3' end against early degradation. The poly A tail also has an important role in maintaining the mRNA's stability (Humphreys, Westman et al. 2005).

The next process is the splicing of introns. Introns generally start with 5'-GU and ends with AG-3' (Russell 2006). The splicing of introns in the nucleus occurs in spliceosomes which consist of the bound of pre-mRNA to small nuclear

ribonucleoprotein particles (snRNPs). snRNPs are snRNAs associated with proteins. There are five snRNAs involved in the splicing processed. They are U1, U2, U4, U5, and U6 (Aronova, Bacikova et al. 2007). Each of this snRNAs interacts with some proteins to form the snRNPs. There are at least 10^5 copies of snRNPs per cell in the nucleus. The steps of the splicing are the following:

1. U1 snRNP binds to 5' splice junction of the intron.
2. U2 snRNP binds to the branch point sequence located upstream of the 3' splice junction.
3. U4/U6 snRNP and U5 snRNP interact and binds to the U1 and U2 snRNP. This process caused the intron to loop bringing its two ends close to each other.
4. U4 snRNP detach, yields the formation of active splicosome.
5. Inside the splicosome, the snRNPs cleave the intron from the exon at the 5' splice junction and this free 5' intron bind to an A nucleotide in the branch point sequence. This process form a structure called RNA lariat structure.
6. Last step is the excision of the 3' splice junction and then the exon separated by this intron is joined together. The snRNPs are released at this time and the process repeated for each intron.

2.1.2 Translation

The translation is a process of protein synthesis which takes place in the cytoplasm where ribosome is located. The mRNA is translated from the 5' to the 3' direction and the polypeptide (a molecule consist of several amino acids) is made in N

terminal to C terminal direction. The protein is an example of a polypeptide. Translation involves not only mRNA but some other non-coding RNA such as rRNA and tRNA (Amort, Wotzel et al. 2007). There are three basic steps of translation: initiation, elongation, and termination. These three basic steps are similar in prokaryotes and eukaryotes.

2.1.2.1 Initiation of Translation

The initiation process involves mRNA, ribosome subunits, aminoacyl tRNA (the tRNA charged with the first amino acid), initiation factors, guanosine triphosphate (GTP), and magnesium ions (Russell 2006). The initiation factors assist the assembly of the complex needed for the initiation. The initiation process in prokaryotes starts with the binding of the 30S ribosomal subunit to the AUG start codon (a set of three adjoined nucleotides that codes for amino acid, start or termination signal) in the mRNA. The ribosomal subunit binds to three initiation factors (IF1, IF2, and IF3), GTP and magnesium ions. The AUG start codon is not sufficient to signal the ribosomal subunit to bind. It needs a ribosome binding site (RBS) which is located upstream to the 5' site of the AUG start codon (Shine and Dalgarno 1975). The mRNA RBS region is also known as the Shine-Dalgarno sequence.

The next step of the initiation is the binding of the fmet-tRNA (initiator tRNA) to the AUG start codon where the 30s ribosomal subunit is bound. AUG specifies a methionine in both eukaryotes and prokaryotes. Thus the polypeptides in both organisms start with the methionine which in many cases is removed later. The fmet-tRNA has an anticodon 5'-CAU-3' to bind to the AUG start codon. Once the fMet-tRNA binds to the

start codon, IF3 is released. Thus, at this time the 30s initiation complex has been formed. The 30s initiation complex consist of mRNA, 30s subunit, fMet-tRNA, IF1, and IF2. Next, the 50s ribosomal subunit binds and leads to the hydrolysis of the GTP and the release of IF1 and IF2. This process forms the *70s initiation complex*. The 70s ribosome has three sites: the aminoacyl (A) site, the peptidyl (P) site, and the exit (E) site. The fMet-tRNA is bound in the P site and the A and E sites are empty. The difference between the initiation process of prokaryotes and eukaryotes are that in eukaryotes, the initiator methionine is not modified and there are no Shine-Dalgarno sequences. The eukaryotes' ribosome uses a different way to find the AUG start codon. First, eukaryotic initiation factor eIF-4F which consist of several proteins, including the cap binding protein (CBP) binds to the cap at the 5' of the mRNA. Next, complex of the 40s ribosomal subunit with the initiator t-RNA (Met-tRNA), several eIF proteins, and GTP bindes together and moves along the mRNA to find the AUG start codon. The AUG start codon is surrounded in a short sequence well known as the Kozak sequence (Kozak 1986). After the 40s ribosomal subunit finds the start codon, it binds to it and the 60s ribosomal subunit binds, and the eIFs is released except the eIF-4F. This process produce the 80s initiation complex with the Met-tRNA bound to the mRNA in the P site of the ribosome. The poly A tail in eukaryotes' mRNA also play an important role in the stimulation of initiation of the translation.

2.1.2.2 Elongation

The elongation in prokaryotes and eukaryotes is similar to each other. In prokaryotes, after the 70s initiation complex is formed, fMet-tRNA is bound to the AUG

start codon in the P site of the ribosome (Russell 2006). The next codon in the mRNA is in the A site of the ribosome. Next the appropriate aminoacyl tRNA binds to the codon in the A site. The amino acid in the tRNA depends on the codon in the mRNA in the A site of the ribosome. This aminoacyl tRNA is brought into the ribosome and bound to the protein elongation factor (EF-Tu) and GTP. Once the aminoacyl bound to the codon in the A site, the GTP is hydrolyzed and EF-Tu-GDP is released. The EF-Tu is recycled. Second elongation factor, EF-Ts, binds to EF-Tu and detaches the GDP. Then the GTP binds to the EF-Tu-EF-Ts complex to make an EF-Tu-GTP complex simultaneously with the release of EF-Ts. An amynoacyl-tRNA then binds to the EF-Tu-GTP complex and this complex is later bound to the A site in the ribosome when the correct codon is found. The ribosome maintains the two aminoacyl tRNA in the P and A sites in the proper positions so that peptide bond can be created between the two amino acids. After the peptide bond is created, a tRNA without an attached amino acid is left in the P site. The t-RNA in the A site is now a peptidyl tRNA because it has more than one amino acid attached to it. The last step of the elongation is translocation. The ribosome moves one codon towards the 3' of the mRNA. In prokaryotes, this process required the protein elongation factor EF-G. An EF-G-GTP complex binds to the ribosome, followed by the GTP hydrolysis and the translocation occur with the displacement of the uncharged tRNA from the P site. In eukaryotes, the translocation process is similar to prokaryotes, except that the elongation factor is eEF-2 which functions like the EF-G does (Nilsson and Nygard 1984). The uncharged t-RNA moved from the P site to the E site, blocking the next aminoacyl tRNA from binding to the A site until the translocation is complete and the peptidyl tRNA is bound appropriately in the P site. After translocation is done, A site

is empty and aminoacyl tRNA with the correct anticodon bind to the codon in this new empty A site. This whole process continues until the process reaches the stop codon (Russell 2006).

2.1.2.3 Termination

The termination factors or release factors (RFs) assist the ribosome in recognizing the stop codon (Bertram, Innes et al. 2001). Prokaryotes have three RFs (RF1, RF2, and RF3). RF1 recognizes UAA and UAG, and RF2 recognizes UAA and UGA. RF3 doesn't recognize any of the stop codons but it stimulates the termination events. Meanwhile, in eukaryotes, a single eukaryotic release factor (eRF1) recognizes all stop codons and eRF3 stimulates the termination events (Russell 2006). The termination events consist of three main steps:

1. The release of the polypeptide from the tRNA in the P site with the peptidyl transferase as catalyze.
2. The release of the tRNA from the ribosome
3. The dissociation of the two ribosomal subunits and the RFs from the mRNA.

2.2 Non-Coding RNA

Non-coding RNA (ncRNA) is RNA molecule that is not translated into the protein. Non-coding RNA genes produce some important functional molecules and the molecules are very diverse in all cell types (Olivas, Muhlrad et al. 1997). It is known that most of the genome of mammals and other complex organisms are transcribed into ncRNAs. Some of these ncRNAs are processed into some smaller products such as

microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs). There are more small regulatory RNA classes waiting for discovery (Russell 2006). These small regulatory RNAs can control gene expression in physiology and development, which includes chromatin architecture, transcription, RNA splicing, editing, and translation (Hutvagner, Simard et al. 2004). They also play a significant role in disease and genetic variation between and within species (Mattick and Makunin 2006). The classification of non-coding RNA family has been done and the result is stored in Rfam database (Griffiths-Jones, Moxon et al. 2005). According to Rfam, there are four ncRNAs which are found in all three kingdoms of life (Archaea, Bacteria, Eukaryota); they are tRNA, rRNA, RNaseP (tRNA maturation), and SRP RNA (protein export). These four ncRNAs are more understood compared to other family such as miRNA which is found only in the Eukaryota or snoRNA which is found both in Eukaryota and Archaea (Russell 2006).

2.2.1 Transcription of Non-Coding RNA

The transcription of non-coding RNA is different from the mRNA. Because of the limitation of information for some ncRNA families, the transcription describes in this section is the transcription of the rRNA and tRNA (the more understood ncRNA). In the genome, the DNA that has the genes for rRNA are called rRNA transcription units. In these rRNA transcription units, there are some tRNA genes (Russell 2006). For example, in the E. coli, the rRNA transcription units are called *rrn* region (Erdei, Boros et al. 1983). Each rrn contains 16S, 23S, and 5S rRNA coding sequences (Erdei, Boros et al. 1983). There are one or two tRNA genes in the spacer region between the 16S and 23S

and another one or two tRNA genes in the region between the 5S and the terminator (Figure 2.4).

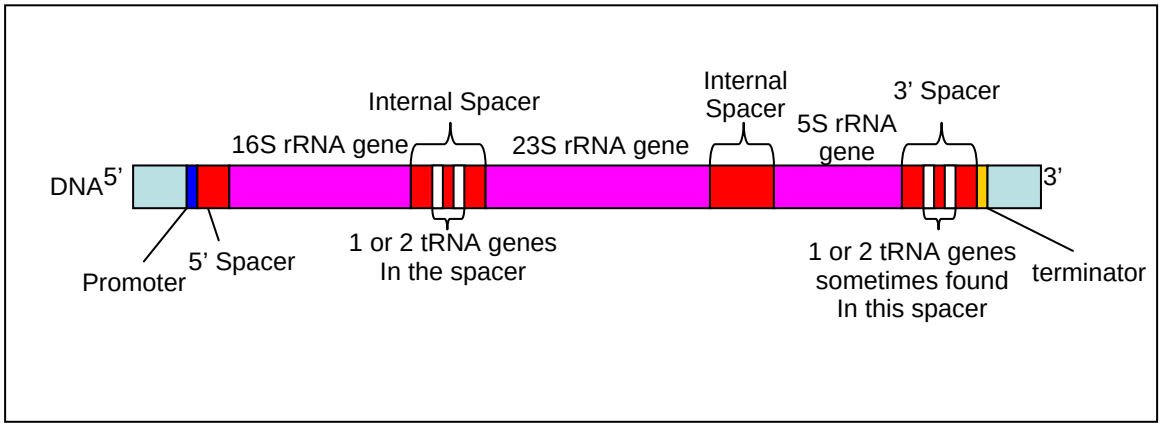

Figure 2.4 E. coli rRNA transcription unit

The result of the transcription of the rRNA units by the RNA polymerase is a single precursor rRNA (pre-rRNA). In the E. coli, the transcription result of the rRNA transcription unit is the 30S pre-rRNA which has 5' spacer, 16S, 23S, and 5S rRNA sequences (separated by spacer) and 3' spacer (Figure 2.6).

2.2.2 The functions of Non-Coding RNA

The non-coding RNA can be classified into three major groups: rRNA, tRNA and mRNA. The non-coding RNA other than rRNA and tRNA were so complicated, non abundant and not stable (Eddy 2001). Several non mRNA other than rRNA and tRNA have been detected, one of which is uridine (U)-rich (Zieve 1981; Busch, Reddy et al. 1982). Following this finding, more small ncRNAs have been isolated using biochemical methods (Eliceiri 1999). Novel non-coding RNA recently discovered is the microRNA (miRNA). The discovery of miRNA was initiated by the identification of lin-4 regulatory

RNA in *C.elegans.* The product of lin-4 is 22 nucleotides RNA processed from a stem loop structured RNA with 61 nucleotides (Horvitz and Sulston 1980). Lin-4 remains strange until the discovery of let-7 which is known as a 21 nucleotides and possibly targets the mRNA of lin-41 and lin-42 genes (Reinhart, Slack et al. 2000). Let-7 is known to be conserved and expressed as 21 nucleotides long in bilaterally symmetrical animal including human, mouse, chicken, polychaete worms, and flies, but not in jelly fish and poriferans (Pasquinelli, Reinhart et al. 2000). The function of this homologous let-7 in the organisms mention above is unknown, but since they have similar behavior with the one in the worms. The assumption was that they function also as a post transcriptional regulation on developmental genes (Eddy 2001). Pasquinelli et al. (2000) named this kind of RNA as small temporal RNA (stRNA)(Pasquinelli, Reinhart et al. 2000). Along with the finding of the lin-4 and let-7, the phenomenon of double stranded RNA interference (RNAi) was under observation. Double stranded RNA was injected to nematodes. This led to a rapid degradation of the homologous mRNA and the loss of function of the phenotype associated with the mRNA (Olsen and Ambros 1999). In plants, RNAi can also cause post transcriptional gene silencing (Carthew 2001; Vance and Vaucheret 2001). This injected RNAi is later cleaved by the identified putative processing nuclease Dicer into small interfering RNA (siRNA) which is about 21-25 nucleotides long (Hutvagner, McLachlan et al. 2001). The function of the RNAi is then suggested as a primitive immune system against RNA viruses and retrotransposons (Sharp 2001; Vance and Vaucheret 2001).

The search of such small RNAs continues and Thomas Tuschl found that Drosophila has endogenous 21 and 22-mers. He also suggest that this might be a natural

occurrence of siRNA in Drosophila (Elbashir, Lendeckel et al. 2001). The fact that RNA with small number of nucleotides such as lin-4 and let-7 belong to nematode, fly, human, and possibly other organisms has later became the foundation of the discovery of microRNA (miRNA). About 55 novel miRNA in *C. elegans* have been identified later on (Lau, Lim et al. 2001). Since then, the identification of novel non-coding RNAs and their functions has been one of a major research in science and this problem has become more and more challenging. Some new discovered functions are control of chromosomes, splicing, RNA editing, translational inhibition and mRNA destruction (Mattick and Makunin 2006).

2.3 MicroRNA (miRNA) biogenesis and functions

MiRNA is a small evolutionary conserved (~22 nucleotides) RNA. They act as a determinant of the destruction or repression of the mRNA target (Nelson, Kiriakidou et al. 2003). The biogenesis of the miRNA starts by the transcription of the endogenous miRNA gene by the RNA Polymerase II. miRNA genes has a minimal of 600 base pairs promoter that is dependent on the RNA Polymerase II (Lee, Kim et al. 2004). Although this promoter is able to undergo a Polymerase II dependent transcription, it lacks of the essential element usually required for initiation of transcription. The process of the transcription leads to the generation of primary miRNA (pri-miRNA) which has a hairpin shaped secondary structures and further process in the nucleus (Gregory and Shiekhattar 2005). The processing of the pri-miRNA is initiated by Drosha which is a product of RNase III endonuclease (Lee, Ahn et al. 2003). Drosha cleaves the hairpin stem on both sides near the primary stem loop and indeed leaving the ~60-70 precursor miRNA (pre-miRNA). The pre-miRNA has 5' phosphate and 3' two nucleotides overhang (Gregory

and Shiekhattar 2005). The pre-miRNA then transported to the cytoplasm by Exportin-5. Once in cytoplasm, Dicer's PAZ domain (another product of RNase III endonuclease) recognized the 2 nucleotides overhang in the 3'. Dicer cleaves the stem of the pre-miRNA hairpin resulting in ~22 nucleotides double stranded RNA (Zhang, Kolb et al. 2004). This double stranded small RNA is the mature miRNA duplex. One strand of the mature miRNA duplex is incorporated with the RNA induced silencing complex (RISC) to be able to affect the expression of the target genes.

MiRNA direct the cleavage of the phosphodiester bond in the target mRNA when perfectly paired to it. The RISC has the "Slicer" activity to conduct the cleavage (Pillai 2005). The study of the inhibition of protein synthesis by miRNA conducted in *C. elegans* with the miRNA lin-4 and the target of this miRNA which is *lin*-14 (Lee, Feinbaum et al. 1993). Lin-4 paired with lin-14 mRNA complementary incomplete in several locations in the 3' UTR and prevented the synthesis of the protein (Wightman, Ha et al. 1993). Similar effect on protein synthesis has been discovered in several other organisms including worms, flies, and humans (Reinhart, Slack et al. 2000; Brennecke, Hipfner et al. 2003; Zeng, Yi et al. 2003). These studies confirm that the complementarity to the target sequences is the key for the outcome of the mRNA target and there is no major change in the mRNA levels although protein synthesis has been affected (Pillai 2005).

In worms, it has been found that the repressed lin-14 (target of the lin-4) is associated with polysomes (Olsen and Ambros 1999). A similar result has been found in lin-28 which is another target of lin-4 (Seggerson, Tang et al. 2002). These results led to a proposal that the translation might be repressed at a step after the initiation. However

the proposal has not been proven experimentally. It has been shown that mammalian miRNA is associated with polirybosome (Krichevsky, King et al. 2003; Kim, Krichevsky et al. 2004; Nelson, Hatzigeorgiou et al. 2004) . It is suggested that the regulation of the miRNA action in the mammals is similar to that in worms. But later studies has proved that the translational mechanism of the target mRNA in worms is different from humans (Pillai 2005).

The thought that animal miRNA incomplete complementary pairing with the mRNA target will not affect the mRNA level has been questioned. Lim et al have found that the mammalian miRNA can affect the mRNA transcripts in the cell (Lim, Lau et al. 2005). Another study found that lin-4 and let-7 miRNA in *C.elegans* act by affecting the stability of the target mRNA (Bagga, Bracht et al. 2005). Lin-4 was claimed not to affect the mRNA level (Olsen and Ambros 1999). These new discoveries have raised a new question if there is translational repression, mRNA degradation, or both. In worms, miRNA lin-4 and let-7 mRNA target degradation is responsible for the reduction of protein production (Bagga, Bracht et al. 2005). However, in *C.elegans* it has been proven by in vitro experiment that mRNA lin-41 with the let-7 complementary site are not cleaved by the human let-7 designed RISC activity (Hutvagner and Zamore 2002). While in mammals, the mammalian endogenous mRNA regulated by miRNA in the translational level without a change in the target mRNA level (Poy, Eliasson et al. 2004; Cimmino, Calin et al. 2005). These results can produce a view that translational repression is the primary event, and the reduction of the target mRNA levels is the consequence of the translational repression (Pillai 2005).

2.3.1 MiRNA and Cancer

There are at least two ways of how miRNA biogenesis affects cancer. First, there has been an indication that the miRNA is associated with some types of cancer such as leukemia (Calin, Dumitru et al. 2002), colon cancer (Michael, O'Connor et al. 2003), lung cancer (Takamizawa, Konishi et al. 2004), and lymphatic cancer (Metzler, Wilda et al. 2004). MiRNA genes even often found in fragile sites and genomic region involved in cancer (Calin, Sevignani et al. 2004). Second, exploitation of the therapeutics of RNA can be done by the examination of how endogenous miRNA apply their regulatory function (Stevenson 2004; Ryther, Flynt et al. 2005). Chronic lymphocytic leukemia (CLL) is the most common form of leukemia in adults which can be credited to a deletion in chromosomal loci 13q14 in more than 50% of the cases (Gregory and Shiekhattar 2005). Some other type of cancers has also been proved to have a connection with the deletion of the 13q14 in various degrees. MiRNA genes *mir-15 and mir-16* is the two miRNA genes maps to the deleted region above. It is known that in more than 68% of CLL cancer patient and most of the examined prostate cancer cell, these two miRNA genes has been deleted or down regulated. In CLL cell, accumulation of the pre-mir-15 has been detected by Northern Blotting leading to a potential shortage of the mir-15 processing (Calin, Dumitru et al. 2002). The regulation of the mir-143 and mir-145 also identified in the colorectal neoplasia (Michael, O'Connor et al. 2003). In lung cancer, the expression of the let-7 miRNA is significantly reduced (Takamizawa, Konishi et al. 2004). The accumulation of the pre-mir-155 has also been found in the Burkitt lymphoma (Metzler, Wilda et al. 2004). The variation of expression of miRNA in the cancer cell leads to a speculation that miRNA usually located in the genomic region involved in

cancer and miRNA may act as both cancer suppressor and an oncogene (Gregory and Shiekhattar 2005).

2.4 The Role of Computational Biology

Computational biology has enhance the development of some research areas, such as sequence analysis, genome annotation, evolutionary biology, pathway prediction, protein structure prediction, and RNA structure prediction (Notredame 2002; Silverman 2003; Edgar and Batzoglou 2006) and RNA secondary structure prediction (Silverman 2003). Most of the prediction algorithm relies on the achievement of the minimum equilibrium free energy. According to basic thermodynamics, the correct RNA fold gives the lowest energy and should be the most stable. The minimum free energy is determined by summing the base pairs, loops, and others. The values of each individual secondary structure element is determined by melting studies with short oligonucleotides (Freier, Kierzek et al. 1986). For the basepairs, simple individual nearest neighbor technique is used (Xia, SantaLucia et al. 1998). This technique assumes that the stability of a base pair is dependent on the adjacent bases. Thus, both basepairing and basestacking technique is used to calculate the total free energy (Silverman 2003).

The combinatorial folding algorithm was introduced in 1982 by Dumas and Ninio. The algorithm determines the minimum free energy of the folding by exhaustively search for the best combination of the folding (Dumas and Ninio 1982). Obviously this kind of algorithm is extremely time consuming and memory inefficient. The time consumption of the algorithm grows exponentially with the size of the input, thus only RNA with small number of nucleotides can be folded with this algorithm. In 1989, Zuker

introduce the recursive algorithm. This algorithm initially calculates the minimum energy of smaller RNA fragments, possibly around 5 nucleotides, and recursively do this for the entire RNA sequence input. The minimum energy value is stored in a matrix, and after the matrix is filled, the algorithm traces back the best alignment (Zuker 1989). This algorithm is faster than the combinatorial. The time is the cubed of the input size, but it can only predict one optimized structure. Successful recursive algorithm that can give several optimized structure from an input is finally developed when Zuker introduced the idea of changing the point of origin when predicting the structure of a circular RNA. The same idea was generalized to be able to solve the secondary structure prediction recursively in linear RNA (Zuker 1989).

RNA folding can be achieved based on conservation among species. This conservation has been used to determine the phylogenic relationships between organisms (Kumar and Rzhetsky 1996). Therefore phylogenic relation between organisms can be used to predict the fold of the RNA. Most of the algorithm using phylogenic comparative analysis depends on the nucleotides sequence alignment. This method can be successful in finding the conserve region of the RNA sequence but it is not globally accurate since no minimum folding energy is calculated. The more sequence in the input, the more accurate the result is, but the less conserve regions are detected. The fact that this algorithm does not take minimum folding energy into account makes the result lacks of biological significance (Ritchie, Legendre et al. 2007)

Genetic algorithm was then introduced by Chen et al. in 2000. This is an algorithm using a comparative method, but the approach is very different from the phylogenic comparative method. This method does not require sequence alignment. Both

structural energy and similarity are considered. The calculation of the free minimum energy is by the nearest neighbor method. This method has been very successful in predicting tRNA and small rRNA (Chen, Le et al. 2000). The limitation of this method is that it can only be used for phylogenicly related sequences. The computational time of this algorithm is worse (Silverman 2003) compared to some of the most recent algorithm like MFOLD.

The MFOLD algorithm is then introduced in 1994 and it relies on the minimum free energy to determine the best structures. The algorithm is recursive and able to give a suboptimal structure (Walter, Turner et al. 1994). MFOLD uses the most accurate thermodynamic data for the calculation of the minimum free energy and the detail algorithm has been used to develop the more accurate algorithm for RNA structure prediction (Mathews, Sabina et al. 1999). More sophisticated and efficient algorithms have been developed to predict RNA secondary structure faster and more accurate (Silverman 2003). This algorithm combines the use of comparative sequence analysis and minimum energy calculation. It uses the thermodynamics parameter used by MFOLD. It involves a complicated dynamic programming with 4 dimensional dynamic programming matrix (Mathews and Turner 2002). The result of this algorithm is better than just using minimum energy calculation. It can predict approximately 86% for known basepairs compare to only 50-60% by using only minimum energy calculation (Silverman 2003). The limitation of this algorithm is yet again the dependency of its accuracy on the accuracy of the experimental data used for the thermodynamics parameter. Another limitation is that this algorithm can not predict the structure of more than two common sequences due to the use of 4 dimensional dynamic programming matrix for two common

sequences. The increase in the number of common sequences significantly increases the number of the dimension of the matrix. There are still rooms for improvement of the RNA secondary structure prediction algorithm either in the efficiency of the algorithm or the accuracy of the algorithm. This can be achieved by paying more attention to the thermodynamics parameter used in the algorithm and allowing input from experimental data (Silverman 2003).

2.4.1 Multiple Sequence Alignment

Sequence alignment is a way to arrange DNA, RNA, or protein primary sequences in order to be able to see similarity that may be a result of functional, structural or evolutionary relationship between the sequences. Multiple sequence alignment algorithms have been a developing research area since the first program to do multiple sequence alignment (MSA) (Edgar and Batzoglou 2006). The basic idea of the MSA is to find the most similar sequence by giving a score to each match, mismatch and gaps. The most naïve approach is to try all possible combination of the arrangement until the maximum score is obtained. Although this technique gives the most accurate result, it is very impractical since it requires a huge amount of computational time and space (Notredame 2002; Edgar and Batzoglou 2006). The time required to try all possible combination of the arrangement grows exponentially with respect to the input sequence size (Wang 1994). Thus, some approximation and heuristic technique is being develop to reduce the time and space needed for the computation (Feng and Doolittle 1987).

A technique of "progressive alignment" is the most widely used technique to solve multiple sequence alignment (Edgar and Batzoglou 2006). The algorithm aligns N

sequences by performing N-1 pairwise alignment. Based on the idea that the sequences are phylogenetically related, the relationship between these input sequences is used to guide the alignment. The most influential component of this algorithm is the scoring scheme of the pairwise alignment. The scoring scheme can be divided into two groups which are matrix based and consistency based (Notredame 2002). The matrix based scoring scheme which is used by some well known program such as CLUSTALW (Thompson, Higgins et al. 1994), and Multiple Sequence Comparison by Log Expectation (MUSCLE) (Edgar 2004). This algorithm use substitution matrix to score the matching of the arranged letter. Whereas, the consistency based scoring scheme considered more information to assess the score for the arrangement, it input a collection of global and local alignment in a primary library, and thereafter the algorithm uses this primary library as a position specific substitution matrix for the progressive alignment. The goal of the algorithm is to have a multiple sequence alignment which is consistent to the alignment contained in the primary library. The algorithm has been used to develop several programs such as T-Coffee (Notredame, Higgins et al. 2000), and Profile Consistency Multiple sequence Alignment (PCMA) (Pei, Sadreyev et al. 2003) which has a better time requirements than T-Coffee. There are some other programs developed with consistency based scoring scheme. Probabilistic Consistency (ProbCons) (Do, Mahabhashyam et al. 2005) utilize the Bayesian probabilistic consistency scoring scheme and pair hidden markov model to fill out the primary library while the substitution cost calculated by the Bayesian statistics. Multiple sequence Alignment improved by using hidden Markov Models with Local Structural information (MUMMALS) (Pei and

Grishin 2006) which combine the ProbCons method and PCMA method with the addition of the local structure information to do the alignment.

Consistency method is more accurate than the matrix based scoring method although the time requirement for the computation of consistency method is N (the number of the input sequences) times higher than the matrix based method (Blackshields G 2006). The fact that the accuracy of these available alignment programs is similar makes it more difficult than ever to objectively choose the better program among them. A consensus method M-Coffee can be used to solve this problem (Wallace, O'Sullivan et al. 2006). M-Coffee was created based on T-Coffee. The input is a sequence data set and the algorithm creates the primary library by using several MSA methods. It uses T-Coffee to finally compute the most consistent alignment. When used with the most accurate available MSA programs, M-Coffee can produce 67% of the time a better MSA compared to ProbCons (Wallace, O'Sullivan et al. 2006). Although M-Coffee performs better than the other, the limitation is that it does not work nicely with remotely homologous sequences. Thus, more improvement on MSA depend only on primary sequence is very difficult (Notredame 2002).

The use of structural and homology data can be crucial to the progress of MSA. Structural extension was introduced by Taylor for protein alignment (Taylor 1986) and template based algorithm has been one of many algorithms using this approach since then (Armougom, Moretti et al. 2006). The principle of the structural extension technique is to use BLAST to identify the structural template in the protein data bank for each of the sequence, aligning the template, and map the original sequences to their template's alignment (Notredame 2007). The result of the alignment with the templates are then

stored in the primary library and used with the consistency method to compute the final MSA.

Besides the use of structural information, homology has also been introduced. DbClustal Package is an example of programs that use homology extension technique (Thompson, Plewniak et al. 2000). This technique is using a profile rather than structure. PSI BLAST (Notredame 2007) is used to profile each sequence and this profile is utilized to build templates for the alignment. This is possible because the profiling contains evolutionary relationship information between each sequence. The fact that MSA is the fundamental tools to conduct further biological research leads to the continuous search of better algorithm to solve this particular problem (Notredame 2002).

2.4.2 MiRNA Prediction Algorithm

Most of the miRNA prediction algorithm use sequence and structure alignment as the fundamental technique. BLASTN (Altschul SF 1990) is a tools to search for DNA similarity. The input is a DNA sequence and the output is the most similar DNA sequence from the database that the user specifies (http://www.ncbi.nlm.nih.gov/BLAST/). BLASTN was also used as a tool to predict miRNA by sequence similarity. The known miRNA sequence used as the input of the BLASTN and the output from the program will be the new miRNA candidates (Pasquinelli, Reinhart et al. 2000). This approach is the trigger of the creation of more sophisticated algorithms for predicting novel miRNA. MFOLD is a tool that can predict a secondary structure from a primary RNA sequence input. The program uses the minimum energy of the secondary structure as a parameter of the prediction (Zuker 1994).

Most of the miRNA identified exclusively only by molecular biology experiment until the year of 2003 (Lim, Lau et al. 2003). This was due to the limited computational tools available for miRNA prediction (Thomassen, Rosok et al. 2006). There are three major factors that made computational miRNA prediction possible. First, miRNA derived from the pre-miRNA which has a stem loop (hairpin) secondary structure with ~70-100 nucleotides. Second, miRNA is conserved between genome of related organisms. Third, miRNA has a unique characteristics of evolutionary divergence (Lai, Tomancak et al. 2003).

In 2001, Lee and Ambros used both bioinformatics and cDNA cloning technique to identify novel miRNA in *C.elegans*. They found the conserved sequence of *C.elegans* in *C.briggsae* that has a pre-miRNA secondary structure and features similar to lin-4 and let-7. They uses the MFOLD program to identify the folding of the primary conserved sequence (Zuker 1994). They reported 15 novel miRNA out of which two were the result of computational approach (Lee and Ambros 2001). Another computational approach is a program MiRScan (Lim, Lau et al. 2003). MiRScan was created to identify miRNA genes conserve between genome and was first applied to *C.elegans* and *C.briggsae.* This program used together with sequencing clone and was able to detect 30 additional new miRNA. The program starts by scanning genome A and B. It scans genome A for the sequence which can fold into a hairpin secondary structure and check if this sequence exists and conserved in genome B. This step finds the homologous pre-miRNA in both genomes. The program utilizes experimentally verified miRNA as a training data and computes a score for all the initially recognized sequences. This program was able to detect 35 novel miRNA in *C.elegans* where 16 of them were experimentally verified. The

program uses a threshold of half detection of the known miRNA. With this threshold, in the worst case scenario, the program had 0.7 sensitivity. The limitation of this program is that it relies on the homology. This fact limits the ability of the program to detect miRNA of organisms with far phylogenic distant. Also if the miRNA is not conserved between species, this program is not a good program to predict such type of miRNA.

In 2003, a program miRSeeker was introduced. This program is a computational prediction of miRNA in Drosophila (Lai, Tomancak et al. 2003). The approach is to first search the sequences in *D. melanogaster* and *D. pseudoobscura* for a transcript that can form a hairpin secondary structure and has a pattern of nucleotide divergence of the known miRNAs. At first Lai et al (2003) aligned 24 pre-miRNA sequences from *D. melanogaster* and *D. pseudoobscura* and found that they are more conserved compared to the protein coding region. The algorithm consists of three major steps: (1) align the intronic and intergenic region of the *D.melanogaster* and *D.pseudoobscura.* (2) slides a window along the conserved region and use MFOLD to determine the secondary structure minimum free energy that is potentially formed by this region. A minimum of 23 nucleotides arm length is required along with at most -23 kcal/mol energy, for one isolated pre-miRNA arm. MFOLD also folds the 2 DNA strands inside the sliding window. (3) fit the pre-miRNA candidates to the pattern of the stem loops of the initial 24 training sets. This procedure yielded 208 new pre-miRNA candidates which include 18 from the training set being in the top 124 candidates with the highest score. They also estimate that miRNA makes up 1% of the Drosophila genome. They claimed that this algorithm excludes the detection of at least one known miRNA which is mir-100. This method is limited by only able to predict miRNA of closely related organisms.

A phylogenic approach has also been used to identify novel miRNA in human (Berezikov, Guryev et al. 2005). The technique is known as the phylogenetic shadowing (Boffelli, McAuliffe et al. 2003). This method allows the comparison of closely related organisms in the nucleotide level. The approach has resulted in an accurate conserved region identification among some organisms (Brown and Sanseau 2005). Berezikov et al (2005) first compared 100 miRNA regions in 10 different primates. They derived a profile which provided information of the variation in the loop of the hairpin, conservation of the stem of the hairpin, and the less conserve region of the sequences flanking the hairpin. This profile was used to identify a potential novel miRNA by pairwise alignment of more divergent organism such as human and mouse or human and rat. Given that miRNA has lower minimum free energy than a random sequences (Bonnet, Wuyts et al. 2004), the candidates from the pairwise alignment were further filtered according to the minimum free energy value. They are able to identify 976 human miRNA where 80% of them are known miRNA. The combination this approach, database search and the northern blot analysis further identified 200-300 novel miRNAs.

ERPIN is developed by Lambert et al. (2004). This is not specifically made to identify miRNA, but it is a general RNA motif identification program that utilizes RNA sequence alignment as an input and identified related sequences using a dynamic programming algorithm (Lambert, Fontaine et al. 2004). A "dynamic programming" (Bellman 1955) is a faster recursive method of solving problem which has an overlapping subproblems and optimal substructure properties. A problem is said to have an overlapping subproblem if the problem can be broken down into some smaller problems which are reused several times. An optimal substructure means the optimal solution of

the subproblem that can be used to construct the optimal solution for the original problem. ERPIN uses RNA sequence alignment and secondary structure annotation as an input. From this input, the program will create a profile for each strand in the alignment. Dynamic programming is then used to identify significance occurrences of these profiles in the database sequences, The profile will create weighted matrix describing the stem and the single strands in the RNA motif (Gautheret and Lambert 2001). The limitation of ERPIN is the fact that some miRNA classes only have a small amount of known miRNA. This will limit the ability to construct the profile.

Another program miRAlign is created by Wang et all in 2005 (Wang, Zhang et al. 2005). Their approach is to use both sequence and structure information to predict novel miRNA candidates. First, they BLAST all known pre-miRNA with a sensitive parameter (word length 7 and E-value cutoff 10) against the genome database. The sequence hit by the BLAST is later cut from the genome with 70 nucleotides flanking sequences to each end. These sequences are scanned with a 100 nucleotides sliding window with a step of 10 nucleotides. The sequences overlapped with the repeat sequences are discarded and the rest is treated as the miRNA candidates to be scored. They then use RNAFold (Hofacker, Fontana et al. 1994) to undertake the folding and calculate the minimum free energy of the of the candidates strands and its complementary. Only strands with energy lower than -20 kcal/mol are kept for further investigation. The candidates that passed this step are aligned with the pairwise sequence alignment technique to all the known ~22 nucleotides miRNA. They then check to ascertain that the position of these ~22 nucleotides is not in the loop of the hairpin candidates and they measure the secondary structure conservation by undertaking pairwise structure alignment. Candidates that pass

through all of these steps and scoring scheme are selected as novel miRNA candidates. Wang et al claim that their program outperforms the previously available program for miRNA prediction, however, it should be noted that miRAlign only tested animal data. It was able to detect 59 new miRNA candidates in *Anopheles gambiae*. Later 37 of these were been reported in the miRNA database (Griffiths-Jones 2004; Griffiths-Jones, Grocock et al. 2006).

In 2005, a program miR-abela was introduced (Sewer, Paul et al. 2005). The program takes three major steps: (1) it takes the input sequence and extracts the genomic region which has the same stem loop secondary structure with the input. This is the robust stem loop. (2) for each stem loop, the program will assign a score based on the similarity of this stem loop to the known human pre-miRNA. Since the precise structural feature of the stem loop that contributes to the pre-miRNA recognition by the enzyme during the miRNA biogenesis is unknown, the program does the following process. It uses two training sets, the known human pre-miRNA as a positive data set and a randomly picked subsequence from the genome regions such as tRNA, rRNA, and mRNA genes as a negative data set. From this data set, the program creates a model which describes the relative contribution (weight) of each of the features to the score given to each of the stem loops. The score measures the distance of the candidate pre-miRNA stem loop from the stem loops in the negative data set, (3) a probabilistic framework is developed. This framework allows the program to estimate the pre-miRNA content in the input genomic sequence from the score given to all stem loops identified in this input sequence. This framework is not dependent of arbitrary score cut off and is used to determine the predicted pre-miRNA stem loop.

Another program ProMir is introduced in 2005 (Nam, Shin et al. 2005; Nam, Kim et al. 2006). This program introduced the probabilistic co learning approach based on paired Hidden Markov Model (HMM) to implement general miRNA prediction technique to identify close and distant homologs. It uses both sequential and structural information in probabilistic framework to decide if a miRNA gene and mature miRNA presents in a region. This is done by recognizing the signal of the site cleaved by Drosha during the miRNA biogenesis. The program used positive data and negative data as their training sequences and use RNAFold (Hofacker, Fontana et al. 1994) to predict all the folding of the training sequences. A pre-miRNA can be viewed as a paired sequence. It forms a stem loop structure and the stem is a matched paired sequence. Like in the sequence alignment, there will be a match, mismatch, insertion or deletion in this pre-miRNA. In the sequence alignment, the alignment is a match for same nucleotides. While in the case of pre-miRNA stem loop, the alignment is a match for nucleotides and the complementary, for example and an **A** matches with **U**, a **G** matches with **C**. This program takes the loop of the pre-miRNA as an order of mismatches and insertions. The HMM has a structural state (match, mismatch, insertion, deletion) and hidden state (the information of the mature miRNA region) match state M can emit a symbol AU, UA, GC, CG, UG, or GU. Deletion state D can emit ●A, ●U, ●C, or ●G. Insertion state, **I**, can emit A●, U●, C● or G●. Mismatch state **N** can emit one of the remaining combinations. The hidden state is a true or a false which indicate if a mature miRNA is there or not. Each of the states has their own transition probability and this probability is used to calculate the outcome of the hidden state. To derive this probability, two parameters are needed: the transition probability and emission probability. Meanwhile, the probability of

a state depends on the probability of the previous state. Using this fact and the two parameters, we can calculate the probability that sequence X generated by such probabilistic co-learning model P. The program then calculated the probability of a region being a mature miRNA. They use 22 ± 2 base pairs sliding windows to scan through the input stem loops and compute the maximum probability P starting from 5' and starting from 3' strand of the pre-miRNA. If the value of P is higher than a previously selected threshold, then the given candidate is a pre-miRNA. The limitation of ProMir is that the program was made with human pre-miRNA as their training data set. Thus, miRNA in the other organisms which do not have homology with the human miRNA might not be detected.

CHAPTER 3

Specificity and Sensitivity of PROMIR, ERPIN and MIR-ABELA in predicting pre-microRNAs in the Chicken Genome[1]

[1] Sebastian, B. and S.E. Aggrey. To be submitted to *In Silico Biology*.

3.1 Abstract

MicroRNAs (miRNAs) are endogenous 21 to 23 nucleotide molecules that are known to regulate about 30% of protein-coding genes through the RNA interference pathway. Low level expression of some miRNA has limited their *in-situ* discovery. Such limitations could be ameliorated by *in-silico* methodologies. The efficacies of three major algorithms (ERPIN, ProMir and miR-abela) in detecting miRNA were evaluated using chicken pre-miRNA data. The sensitivity of ProMir and miR-abela were 53% and 57%, respectively. ERPIN has only 25% of the miRNA classes, and within the classes that were available the sensitivity was 93%. All the 3 algorithms did not predict any false positives from 200 negative data. The efficiencies of the existing programs are low for chicken data and an efficient algorithm may be needed to predict novel chicken pre-miRNAs.

3.2 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression by various mechanisms (Bartel, 2004). MicroRNAs are believed to act via the RNA silencing pathway to regulate mRNA stability or translation, or chromatic structure in the same way as small interfering RNAs do in the RNAi pathway (Schwarz, Tomari et al. 2004). In human, it is estimated that up to 30% of the genes may be regulated by such mechanism (Ouellet, Perron et al. 2006). Developmental studies in the nematode *Caenorhabditis elegance* led the discovery of the first miRNA, lin-4 (Ambros and Horvitz 1984; Ambros 1989; Lee, Feinbaum et al. 1993). Experimental results have shown that miRNAs are also associated with apoptosis (Hipfner, Weigmann et al. 2002;

Jovanovic and Hengartner 2006), maintenance of heterochromatin (Reinhart and Bartel 2002), hematopoietic cell differentiation (Chen, Li et al. 2004), oncogenesis (Michael, O'Connor et al. 2003; Metzler, Wilda et al. 2004; He, Thomson et al. 2005) and the control of viral assembly (Hariharan, Scaria et al. 2005; Cullen 2006).

Like other mRNAs, miRNAs are transcribed by RNA polymerase II as primary miRNAs (pri-RNAs) that require subsequent processing to produce a functional mature miRNA (Donker, Mouillet et al. 2007; Lee, Li et al. 2007; Wang, Medvid et al. 2007). In animals, Drosha, a nuclear RNase III acting with its double stranded RNA binding partner protein DGCR8 (in vertebrates) or Pasha (in invertebrates) cleaves the flanks of the pri-miRNA to liberate an appromixately 70-nucleotide stem loop, the precursor miRNA (pre-miRNA) (Lee, Ahn et al. 2003; Landthaler, Yalcin et al. 2004; Wang, Medvid et al. 2007). The resulting pre-miRNA have 5' phosphate and 3' hydroxyl termini, and two- or three- nucleotide 3' single-stranded overhanging ends, which are characteristics of the Drosha enzyme (Lee, Ahn et al. 2003; Landthaler, Yalcin et al. 2004; Wang, Medvid et al. 2007). Exportin 5, a carrier protein and RanGTP transports the pre-miRNA from the nucleus to the cytoplasm by recognizing the characteristic end structure of pre-miRNA (Lund, Guttinger et al. 2004; Zeng and Cullen 2004). In the cytoplasm, a second RNase III, Dicer makes a pair of cuts at a certain distance from the overhang created by Drosha, thereby defining the other end of the mature miRNAs, liberating an ~21 nts RNA duplex with 2-nts 3' overhangs (Hutvagner, McLachlan et al. 2001; Ketting, Fischer et al. 2001). The first processing step therefore determines the mature miRNA. The miRNA duplex has a structure similar to double stranded small interfering RNAs (siRNAs), except that the mature miRNA is only partially paired to the

miRNA* (RNA at the side of the pre-miRNA stem opposite the miRNA) because the stems of pre-miRNA are imperfectly double stranded (Ritchie, Legendre et al. 2007). The mature miRNA enters the RNA-induced silencing complex (RISC) that represses target gene expression, whereas the miRNA* is degraded (Chendrimada, Finn et al. 2007). The thermodynamic stability of the miRNA/miRNA* duplex determines which strand is loaded into the RISC, which is usually the strand whose 5' end is less stably paired (Khvorova, Reynolds et al. 2003; Schwarz, Hutvagner et al. 2003). The thermodynamic differences arises, in part because of miRNA/miRNA* duplexes contain mismatches and bulges that favor the mature miRNA (Du and Zamore 2005).

Laboratory identification of miRNA has been accomplished (Wang, Zhou et al. 2004; Samols, Hu et al. 2005). However, such procedures do not detect miRNAs expressed at low levels. Computational approaches can be developed to overcome at least, partially these problems. Several computational methods have been have been developed to predict pre-miRNA and find close homologs among related miRNA (Lai et al., 2003; Legendre et al. 2005, etc). According to Legendre et al. (2005), a profile based method provides better miRNA prediction than sequence similarity searches, and can also predict close homologs in animal genomes. Accordingly, the Easy RNA Profile Identification (ERPIN), based on the principle of log-score profiles generalized to base-paired regions of secondary RNA structure profiles to was developed to detect RNA motifs (Gautheret and Lambert 2001). ERPIN conduct simultaneous searches for helical profiles (stem) and dynamic programming alignment of single strand profiles (loop). However, a number of miRNAs do not have close homologs in the sequenced genomes available to date (Pfeffer et al., 2004) and as a result genome-specific algorithms like

Mir-Abela have been suggested (Sewer et al., 2005). Mir-Abela was developed to detect human microRNAs using support vector (SVM) classifier that do not rely on external transcript (Sewer et al., 2005). The SMV is trained using stem loops features of both positive and negative miRNA precursors. Nam et al. (2005) also developed a probabilistic co-learning method based on paired hidden Markov model (HMM), ProMir to identify both close and distant homologs. The ProMir algorithm combines both sequential and structural characteristics of miRNA genes in a probabilistic framework and simultaneously ascertain whether an miRNA gene and a region of mature miRNA are present by detecting the signals for the site cleaved by Drosha.

Identification of miRNA genes has become critical in the understanding of post-transcriptional gene regulation, and the efficacy of computation algorithms predicting these miRNAs need to be evaluated. The objective of this research is to ascertain the sensitivity and specificity of ERPIN, Mir-abela and ProMir in predicting known chicken pre-miRNA.

3.3 Materials and Methods

Data

There are 150 chicken miRNAs in the miRBase (http://microrana.sanger.ac.uk), but only 146 are usable because three of them have not been validated and one is not assigned to any region in the genome. The chicken sequence data from the ENSEMBL genome browser (http://www.ensembl.org/Gallus_gallus/index.html) was used to extend the pre-mirRNA sequences 50 nts upstream and 50 nts downstream (Appendix 1). The 146 pre-miRNA sequences were used as input for the ERPIN, ProMir and Mir-abela

programs. Random sequences of 240 nts from 200 known chicken protein coding genes and tRNAs, rRNA and snRNA genes with 50 nucleotides flanking each sites (Appendix 2) were selected as negative data inputs for ERPIN, ProMir and Mir-abela.

Sensitivity and Specificity

The fidelity of ProMir, ERPIN and Mir-abela using chicken data as input can be assessed by calculating the proportion of both positive and negative data that are correctly identified by any of the programs. Sensitivity was defined as the proportion of the known chicken pre-miRNA that is correctly identified by the algorithm. Specificity is the proportion of the negative data that is correctly identified by the algorithm.

3.4 Results

The sensitivity and specificity of ERPIN, ProMir-g and MiR-abela in detecting known chicken miRNA is shown in Table 3.1. The ProMir-g program can predict 53% of the known chicken pre-miRNAs (Table 3.2), and the success of prediction varied from chromosome to chromosome. About 42% of the known chicken pre-MiRNA are located on the major chromosomes (1-4). Out of the 62 pre-miRNA located on the major chromosomes, ProMir-g can predict 38% of them. The 146 known chicken pre-miRNA are grouped into 88 groups, but only 22 of the groups are available in ERPIN which comprised of 50 pre-mirRNA in the 22 groups. The sensitivity of ERPIN in predicting from the 22 groups was 93% (Table 3.3). Chromosome 4 was an exception where ERPIN predicted 50% of the 8 pre-mirRNA available in its groups.

MiR-abela was able to predict 57% of the known chicken microRNAs (Table 3.4). MiR-abela can predict about 51% of the pre-mirRNA on GGA1-4. The specificity of ProMir-g, ERPIN and Mir-abela was 100% (Table 3.1). They did not provide any false positive results from the negative data. The locations, accession number and identification of the pre-miRNAs and the negative data are presented in Appendices 1 and 2, respectively.

Table 3.1 Sensitivity and Specificity of ProMir[1], ERPIN[2] and MiR-abela[3] in predicting known chicken pre-microRNA.

	ProMir	ERPIN	Mir-abela
Sensitivity	53.42%	93%	57.53%
Specificity	100%	100%	100%

[1] http://cbit.snu.ac.kr/~ProMiR2
[2] http://tagc.univ-mrs.fr/erpin/ (Only 22 of 88 known miRNA classes were available in ERPIN)
[3] http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi

Table 3.2. Prediction of known chicken pre-miRNA using ProMir [(1)]

Chromosome #	# of known pre-miRNA	# of known pre-miRNA predicted by ProMir-g	Percentage Predicted
1	24	13	54.17%
2	17	1	5.88%
3	9	2	22.22%
4	12	8	66.67%
5	1	0	0.00%
6	3	2	66.67%
7	3	0	0.00%
8	4	2	50.00%
9	2	2	100.00%
10	5	2	40.00%
11	2	1	50.00%
12	5	3	60.00%
13	3	2	66.67%
14	2	1	50.00%
15	3	2	66.67%
17	6	5	83.33%
18	1	1	100.00%
19	4	4	100.00%
20	3	3	100.00%
21	4	2	50.00%
23	5	3	60.00%
24	4	4	100.00%
26	6	2	33.33%
27	1	1	100.00%
28	3	2	66.67%
Z	9	8	88.89%
Un_Random	4	2	50.00%
10_random	1	0	0.00%
Not Validated	3		
No Information	1		

[(1)] http://cbit.snu.ac.kr/~ProMiR2

Table 3.3 Prediction of known chicken pre-miRNA using ERPIN [(2)]

Chromosome #	# of known miRNA	# of known miRNA Predicted by ERPIN	# of not available groups	Percentage Predicted out of available miRNA groups
1	24	14	10	100.00%
2	17	5	12	100.00%
3	9	4	5	100.00%
4	12	4	4	50.00%
5	1	1	0	100.00%
6	3	N/A	3	N/A
7	3	1	2	100.00%
8	4	N/A	4	N/A
9	2	2	0	100.00%
10	5	N/A	5	N/A
11	2	1	1	100.00%
12	5	4	1	100.00%
13	3	N/A	3	N/A
14	2	N/A	2	N/A
15	3	N/A	3	N/A
17	6	N/A	6	N/A
18	1	N/A	1	N/A
19	4	N/A	4	N/A
20	3	2	1	100.00%
21	4	1	3	100.00%
23	5	3	2	100.00%
24	4	4	0	100.00%
26	6	4	2	100.00%
27	1	N/A	1	N/A
28	3	1	2	100.00%
Z	9	1	8	100.00%
Un_Random	4	N/A	4	N/A
10_random	1	N/A	1	N/A
Not Validated	3			
No Information	1			

[(2)]http://tagc.univ-mrs.fr/erpin/

Table 3.4 Prediction of known chicken pre-miRNA using miR-abella [(3)]

Chromosome #	# of known miRNA	# of known miRNA Predicted by miR-abela	Percentage Predicted
1	24	13	54.17%
2	17	6	35.29%
3	9	4	44.44%
4	12	9	75.00%
5	1	1	100.00%
6	3	2	66.67%
7	3	2	66.67%
8	4	3	75.00%
9	2	1	50.00%
10	5	4	80.00%
11	2	1	50.00%
12	5	2	40.00%
13	3	3	100.00%
14	2	0	0.00%
15	3	1	33.33%
17	6	5	83.33%
18	1	1	100.00%
19	4	3	75.00%
20	3	3	100.00%
21	4	1	25.00%
23	5	3	60.00%
24	4	4	100.00%
26	6	1	16.67%
27	1	1	100.00%
28	3	2	66.67%
Z	9	5	55.56%
Un_Random	4	2	50.00%
10_random	1	1	100.00%
Not Validated	3		
No Information	1		

[(3)] http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi

3.5 Discussion

The importance of miRNAs in gene regulation has led to the establishment of miRNA database. The miRNA database entries require experimental validation of mature miRNA expression and computational prediction of the corresponding hairpin precursor (He and Hannon 2004). Therefore it is imperative that the computational sensitivity and

specificity of prediction algorithms are high and efficient. Since the number of miRNAs is expected to increase (Legendre, Lambert et al. 2005) it is expected that more novel miRNAs will be discovered in the chicken genome. However, the current algorithms are not suitable for predicting chicken pre-miRNA. Nam et al. (2005) argued that programs based on secondary structures alone cannot predict distant homologs, and as a result the ProMir was developed.

The algorithm of ERPIN is based on profiling the secondary miRNA structure and using dynamic programming to align the loop. For the classes available, ERPIN is every efficient with high sensitivity in predicting pre-miRNAs in chickens. However, it still has a limited application since most of the miRNA classes are not available. Only 25% of the miRNA classes are represented in ERPIN. Secondary structure profiling depends on the fidelity of the training data. Since the sensitivity and specificity of ERPIN are very high, it can be inferred that profiling of secondary structure is an efficient method.

According to Nam et al. (2005) profiling algorithms routinely fail to detect miRNAs that lack detectable homologs. They developed the probabilistic co-learning method as an alternative. The ProMir shows 53% sensitivity and 100% specificity when applied to known chicken pre-miRNA. ProMir has a higher sensitivity (73%) in predicting human miRNA (Nam et al. 2005), but demonstrates a lower sensitivity in predicting chicken pre-miRNA. Therefore, ProMir may not be suitable for predicting chicken pre-miRNA. The ProMir algorithm simultaneously detects the signals for the site cleaved by Drosha, and this method may limit the sensitivity of the program. The program relies on paired Hidden Markov Model (HMM) to scan the stem of the stem loop candidates to determine the signal of the Drosha cleave site. The sites of Drosha

cleavage are determined mainly by the distance from the terminal loop, sequence around the cleavage site, and variability in the stem structure (Zeng and Cullen 2005; Zeng, Yi et al. 2005). Therefore, the inability of the paired HMM to capture most of the factors affecting Drosha cleavage may limit the sensitivity of ProMir.

An *ab initio* method, miR-abela was developed using (Support Vector Machine) SVM to predict miRNAs in the human genome without the use of external transcripts or other genomes, and it retains its value when cross-species conservation plays an important role. The sensitivity of miR-abela in predicting known chicken pre-miRNA was 58% slightly higher than ProMir. The miR-abela algorithm has a reasonable success (43%) in predicting pre-miRNA of *C. elegance* (Sewer et al., 2005). The miR-abela approach is to design a method that can better understand the constraints that define the relation of the pre-miRNA and the related processing enzymes (Sewer, Paul et al. 2005). They are using both positive and negative data as their training data. The limited number of known microRNA in their training data may lead to low specificity. To the contrary more than 5000 negative data used in the training set. This leads to the 100% specificity of miR-abela in predicting chicken pre-miRNA.

The efficacy of ERPIN, ProMir and miR-abela has been examined for chicken data. ERPIN has a limited miRNA classes, but has a very high sensitivity for chicken data. Except for chromosome 4, ERPIN was able to detect pre-miRNAs that both ProMir and miR-abela were not able to detect. It is possible that ERPIN used insufficient training data for miRNA on chromosome 4. The three programs examined in this study are based on different methodologies, but they are able to detect more evolutionary conserved miRNAs. However, they have limited success in detecting chicken pre-miRNA. In

addition, most of the training data for these programs have been from human, mouse and rat. For rapid progress in the discovery of chicken pre-miRNAs, chicken should be included in the training data, or a more efficient chicken specific algorithm should be developed.

References:

Ambros, V. 1989. A Hierarchy of Regulatory Genes Controls a Larva-to-Adult Developmental Switch in C-Elegans. Cell 57:49-57.

Ambros, V., and H.R. Horvitz. 1984. Heterochronic Mutants of the Nematode Caenorhabditis-Elegans. Science 226:409-416.

Chen, C.Z., L. Li, H.F. Lodish, and D.P. Bartel. 2004. MicroRNAs modulate hematopoietic lineage differentiation. Science 303:83-86.

Chendrimada, T.P., K.J. Finn, X.J. Ji, D. Baillat, R.I. Gregory, S.A. Liebhaber, A.E. Pasquinelli, and R. Shiekhattar. 2007. MicroRNA silencing through RISC recruitment of eIF6. Nature 447:823-U1.

Cullen, B.R. 2006. Viruses and microRNAs. Nature Genetics 38:S25-S30.

Donker, R.B., J.F. Mouillet, D.M. Nelson, and Y. Sadovsky. 2007. The expression of Argonaute2 and related microRNA biogenesis proteins in normal and hypoxic trophoblasts. Molecular Human Reproduction 13:273-279.

Du, T.T., and P.D. Zamore. 2005. microPrimer: The biogenesis and function of microRNA. Development 132:4645-4652.

Gautheret, D., and A. Lambert. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. Journal of Molecular Biology 313:1003-1011.

Hariharan, M., V. Scaria, B. Pillai, and S.K. Brahmachari. 2005. Targets for human encoded mircoRNAs in HIV genes. Biochemical and Biophysical Research Communications 337:1214-1218.

He, L., and G.J. Hannon. 2004. Microrna: Small RNAs with a big role in gene regulation. Nature Reviews Genetics 5:522-531.

He, L., J.M. Thomson, M.T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S.W. Lowe, G.J. Hannon, and S.M. Hammond. 2005. A microRNA polycistron as a potential human oncogene. Nature 435:828-833.

Hipfner, D.R., K. Weigmann, and S.M. Cohen. 2002. The bantam gene regulates Drosophila growth. Genetics 161:1527-1537.

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 293:834-838.

Jovanovic, M., and M.O. Hengartner. 2006. miRNAs and apoptosis: RNAs to die for. Oncogene 25:6176-6187.

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 15:2654-2659.

Khvorova, A., A. Reynolds, and S.D. Jayasena. 2003. Functional siRNAs and miRNAs exhibit strand bias (vol 115, pg 209, 2003). Cell 115:505-505.

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. Current Biology 14:2162-2167.

Lee, J., Z.H. Li, R. Brower-Sinning, and B. John. 2007. Regulatory circuit of human microRNA biogenesis. Plos Computational Biology 3:721-732.

Lee, R.C., R.L. Feinbaum, and V. Ambros. 1993. The C-Elegans Heterochronic Gene Lin-4 Encodes Small Rnas with Antisense Complementarity to Lin-14. Cell 75:843-854.

Lee, Y., C. Ahn, J.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 425:415-419.

Legendre, M., A. Lambert, and D. Gautheret. 2005. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics 21:841-845.

Lund, E., S. Guttinger, A. Calado, J.E. Dahlberg, and U. Kutay. 2004. Nuclear export of microRNA precursors. Science 303:95-98.

Metzler, M., M. Wilda, K. Busch, S. Viehmann, and A. Borkhardt. 2004. High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. Genes Chromosomes & Cancer 39:167-169.

Michael, M.Z., S.M. O'Connor, N.G.V. Pellekaan, G.P. Young, and R.J. James. 2003. Reduced accumulation of specific microRNAs in colorectal neoplasia. Molecular Cancer Research 1:882-891.

Ouellet, D.L., M.P. Perron, L.A. Gobeil, P. Plante, and P. Provost. 2006. MicroRNAs in gene regulation: When the smallest governs it all. Journal of Biomedicine and Biotechnology:168.

Reinhart, B.J., and D.P. Bartel. 2002. Small RNAs correspond to centromere heterochromatic repeats. Science 297:1831-1831.

Ritchie, W., M. Legendre, and D. Gautheret. 2007. RNA stem-loops: To be or not to be cleaved by RNAse III. Rna-a Publication of the Rna Society 13:457-462.

Samols, M.A., J.H. Hu, R.L. Skalsky, and R. Renne. 2005. Cloning and identification of a MicroRNA cluster within the latency-associated region of Kaposi's sarcoma-associated herpesvirus. Journal of Virology 79:9301-9305.

Schwarz, D.S., Y. Tomari, and P.D. Zamore. 2004. The RNA-induced silencing complex is a Mg2+-dependent endonuclease. Current Biology 14:787-791.

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

Wang, J.F., H. Zhou, Y.Q. Chen, Q.J. Luo, and L.H. Qu. 2004. Identification of 20 microRNAs from Oryza sativa. Nucleic Acids Research 32:1688-1695.

Wang, Y.M., R. Medvid, C. Melton, R. Jaenisch, and R. Blelloch. 2007. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nature Genetics 39:380-385.

Zeng, Y., and B.R. Cullen. 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Research 32:4776-4785.

Zeng, Y., and B.R. Cullen. 2005. Efficient processing of primary microRNA hairpins by drosha requires flanking nonstructured RNA sequences. Journal of Biological Chemistry 280:27595-27603.

Zeng, Y., R. Yi, and B.R. Cullen. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. Embo Journal 24:138-148.

CHAPTER 4

MiR-Explore: A Computational Approach for Predicting pre-microRNAs in the Chicken Genome [1]

[1] Sebastian, B. and S.E. Aggrey. To be submitted to *In Silico Biology*.

4.1 Abstract

Although several microRNAs (miRNAs) have been identified through wet laboratory experimentation, such approaches can detect only abundantly expressed miRNAs or close homologs and such limitations have been ameliorated by computational methods. Current computational programs are not efficient in predicting chicken pre-miRNA and as a result the miR-Explore program has been developed to predict chicken pre-MiRNA. The program develops consensus secondary structures profile of miRNA classes and aligns query sequences and assigning a binary score to matches and mismatches. The sensitivity of miR-Explore shows 89% sensitivity and 100% specificity in identifying known chicken pre-mirRNA. The 89% sensitivity is an improvement over ProMirII and miR-abela which have 53% and 57% sensitivities, respectively. The efficiency of mir-Explored can improved to scan the chicken genome for novel pre-mirRNAs.

4.2 Introduction

MicroRNAs (miRNAs) are about 21-25 nucleotides long non-coding RNA that are widely conserved among species and may serve as antisense regulators of other RNAs in diverse organisms (Ambrose, 2001; Ruvkun, 2001; Lai, 2003; Bartel and Bartel, 2003; Bartel, 2004). The progenitor miRNAs (pri-miRNAs) are long transcripts consisting of either one or multiple miRNA precursors (pre-miRNA) (Lee, Jeon et al. 2002). The pri-miRNA is cut by the Drosha enzyme in the nucleus into one or more 60-70 nts long pre-miRNA imperfect stem-loop (hairpin) structures (Lee, Ahn et al. 2003). The pre-miRNAs are transported from the nucleus into the cytoplasm by Exportin 5 and

hereafter, cleaved by another enzyme Dicer in an imperfect double stranded duplex (Hutvagner, McLachlan et al. 2001; Ketting, Fischer et al. 2001; Zhang, Kolb et al. 2004). Usually only one side of the stem encodes a mature miRNA of ~22 nts (references), however the process of selecting which stem and region of the pre-miRNA that becomes a mature miRNA is still not fully understood (Thomassen et al. 2006). The RNA double stranded duplex is unwound and the mature miRNA is incorporated into an effector complex, the RNA induced silencing complex (RISC) for transport to its target (Schwarz, Tomari et al. 2004; Chendrimada, Finn et al. 2007; Wang, Medvid et al. 2007). The other strand is rapidly degraded.

Animal miRNAs appears to regulate various biological processes such as development (Darnell, Kaur et al. 2006; Zhao and Srivastava 2007), fat metabolism (Xu, Vernooy et al. 2003; Miska, Alvarez-Saavedra et al. 2004; Telernan and Cohen 2006), apoptosis (Hurt and Farrar 2007; Mott, Kobayashi et al. 2007; Ye, Zhang et al. 2007) and (Debernardi, Skoulakis et al. 2007; Lawrie 2007). In addition, animal miRNAs are capable of being active through imperfect complementation and subsequently has profound implications for determining the spectrum of target genes (Brown and Sanseau 2005). Some miRNAs are differentially expressed in specific tissues (Bloomston, Frankel et al. 2007) or differentiated expression levels in tumor tissues (Mathupala, Mittal et al. 2007). The identification of miRNAs targets is critical in understanding their regulatory function. However, many of the miRNAs are yet to be discovered (Zhang, Pan et al. 2006) because of slow wet laboratory procedures and limitations in the algorithms developed for *in-silico* discovery.

The major mirRNA prediction included BLASTN (Altschul SF 1990; Pasquinelli, Reinhart et al. 2000), MirSeek (Lai, Tomancak et al. 2003), MirScan (Lim, Lau et al. 2003) and Easy RNA profile identification (ERPIN) (Gautheret and Lambert 2001), Mir-abela (Sewer at al., 2005) ProMir (Nam, Shin et al. 2005), ProMir II (Nam, Kim et al. 2006). However, these algorithms could not successfully predict all the known chicken pre-miRNA (Chapter 3 this thesis). The underlying assumption of MirSeek and MirScan is that the stem loop of rare miRNA is conserved in the same pattern as abundant miRNA, however some recently discovered mammalian miRNA genes appear to be less conserved in fish (Bartel and Chen 2004). Most miRNAs are well conserved in secondary structure, and sequence alignment may fail to detect distant ancestors that diverged in sequence but kept conservation in structure (Legendre et al., 2004). Algorithms like ERPIN and PromirII were developed based on both on sequence and structure information, however, construction of profiles from limited family members limit the efficiency of the profile constructed from such families. On the contrary, most non-coding RNAs are also hard to predict *de novo* in large genome scans and also hard to recognize without annotation (Hertel and Stadler, 2006), therefore given a well constructed motif from conversed secondary structure from known families should be able to predict novel pre-miRNA within the family.

Identification of pre-miRNAs that yield mature miRNAs is the most critical step in miRNA gene predictions. Therefore, we present a computational approach named MiR-Explore that detects pre-miRNA based on consensus secondary structure alignment derived from multiple species for miRNA families. The goal is to devise a scoring matrix that identifies nucleotide pairs that fold into the same structure in each pair position.

4.3 Materials and Methods

MiR-Explore

Consensus secondary structure prediction

The primary sequences of all known pre-miRNAs are downloaded from miRBase (http://microrna.sanger.ac.uk/sequences/) (Griffiths-Jones, Grocock et al. 2006).Genome sequences of the pre-miRNAs of up to seven species were used, including *C. elegans* (http://www.ensembl.org/Caenorhabditis_elegans/index.html), *G. gallus* (http://www.genome.wustl.edu/projects/chicken), *H. sapiens* (http://www.ensembl.org/Homo_sapiens/index.html), *X. tropicalis* (http://www.ensembl.org/Xenopus_tropicalis/index.html), *D. rerio* (http://www.ensembl.org/Danio_rerio/index.html), *M. musculus* (http://www.ensembl.org/Mus_musculus/index.html), and *D. melanogaster* (http://www.ensembl.org/Drosophila_melanogaster/index.html), were aligned for each mir class. A consensus secondary structure from multiple alignment of pre-miRNA from the seven species (depending on the availability of the known miRNA class of the species) was predicted by MARNA (http://www.bio.inf.uni-jena.de/Software/MARNA/index.html) (Siebert and Backofen, 2005). A consensus structure sequence comprise of the two arms in the hairpin structure (**I** and **J**) and the loop (**L**). An example of a consensus structure from multiple alignment of let-7 is shown in Figure 4.1.

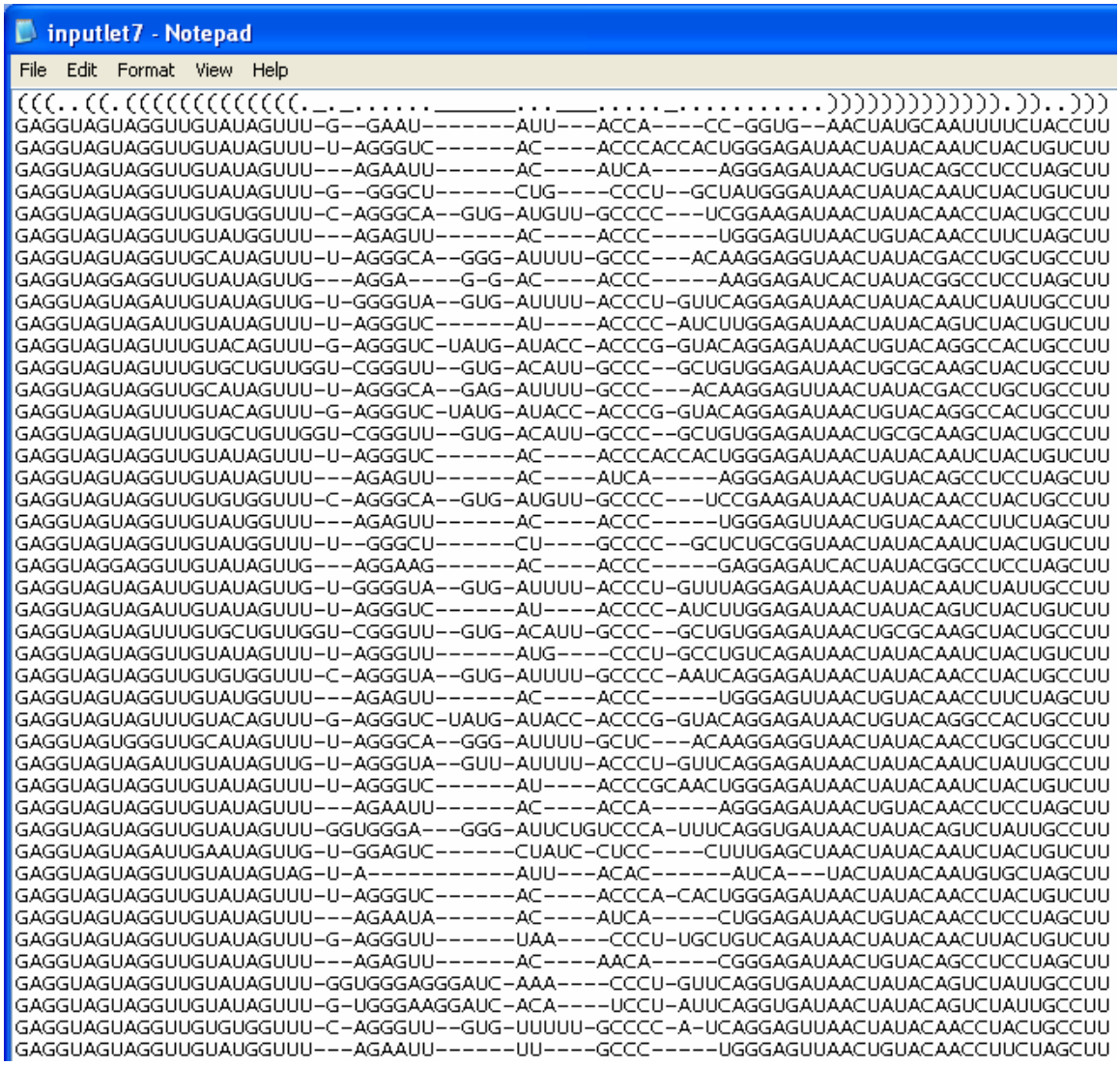

Figure 4.1 Consensus structure from alignment of let-7

Scoring

A 5 x 5 matrix (A, G, C, U and -) is created for each pair in the hairpin structure, **I** and **J** in the consensus structure. All the nucleotide pairs have unique positions, and hereafter will be referred to as pair position. The number of matrices created for each pre-miRNA class will depend on the number of pairs generated from the consensus secondary structure. The default value of all positions in the matrix is set to zero. The

nucleotide pair (I and J) of each of the multiple aligned sequences from which the consensus structure was developed is assigned a value 1. The scoring systems basically identifies the nucleotide pairs that folds into the same structure in each pair position. An example of one score matrices of one stem nucleotide pair of let-7 is shown in Figure 4.2. A window with size corresponding to the maximum number of nucleotides in the training sequences was used to create the consensus structure.

	A	G	C	U	--
A	0	0	0	0	0
G	0	0	1	0	0
C	0	0	0	0	0
U	0	0	0	0	0
--	0	0	0	0	0

Figure 4.2 Scoring matrix of one pair in the let-7.
(In this matrix, in the particular position, the pair is between nucleotide G and C).

Candidate sequence alignment and Iteration

The similarity between a query input sequence and the consensus structure is measured by sliding the window in the query input sequence to scan the 5 x 5 matrices created. A match is scored one, and a mismatch, zero. The maximum score will indicate the best position in which the query sequence is aligned with the consensus structure. The consensus sequence built from multiple sequence alignment and is bound to have gaps in both arms of the hairpin structure, **I**, **J** and the loop, **L**. However, the input query sequence is presented for alignment with no gaps. The gaps in **I**, **J** and **L** of the consensus sequence are eliminated sequentially by sliding the consensus structure over the query sequence and iterated for the best alignment score. The number of iterations performed is equal to product of the total number of gaps in I, J and L. Since the maximum score is

different for secondary structures of different sizes, we define the normality alignment score (**nas**) of the query sequence (qs) and training sequence (ts) as

$$\textbf{nas} = (qs/ts) \times 100$$

The normality alignment score will range from 0 to 100. The higher this value, the more similar are the two structures. It is expected that the **nas** of pre-mirRNA and none pre-mirRNA (negative control sequences) will follow a bimodal distribution. The **nas** of none pre-mirRNA query sequences in each pre-mirRNA class should serve as a cutoff value for that class. A query sequence with **nas** above the cutoff value within a pre-mirRNA class is designated as putative novel pre-mirRNA. An overview description on the Mir-Explore procedure is shown in Figure 4.3. An example of the distribution of the score both from positive and negative data of let-7 are shown in Figure 4.4. In Figure 4.4, the negative data score range from 33.33 to 66.67 and the positive data score range from 77.77 to 100.

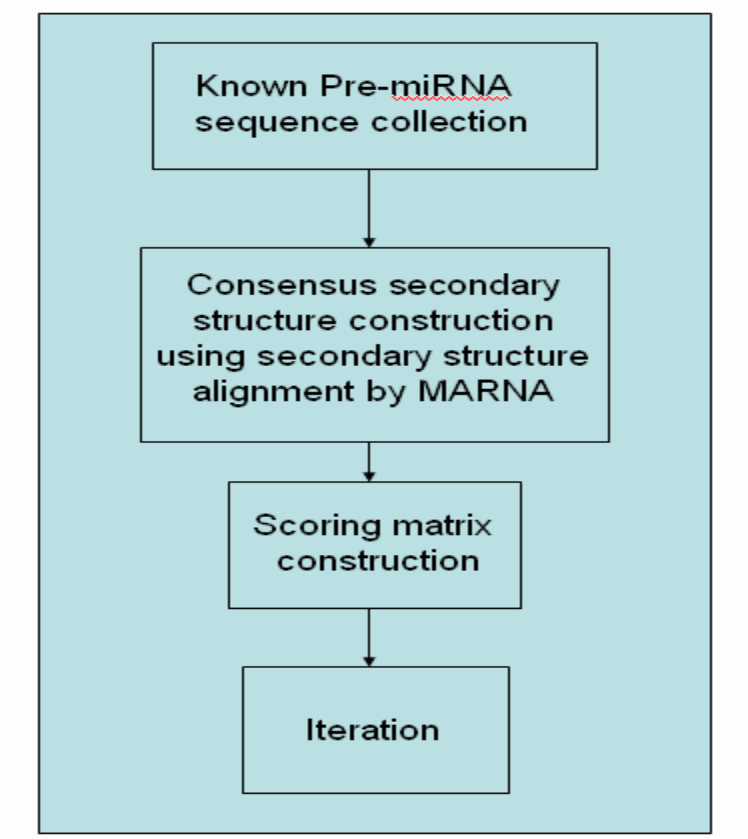

Figure 4.3 Overview of mir-Explore

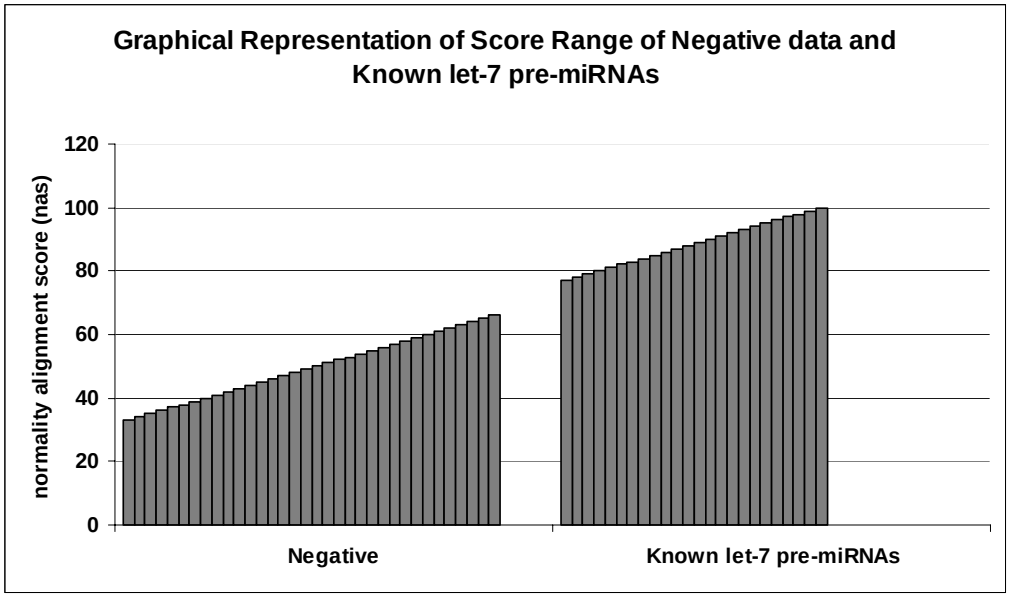

Figure 4.4 Score Range of Negative data and known let-7 pre-miRNAs

Data

One hundered and forty-six chicken miRNAs in the miRBase (http://microrana.sanger.ac.uk) were used as positive data. The chicken sequence data from the ENSEMBL genome browser (http://www.ensembl.org/Gallus_gallus/index.html) was used to extend the pre-mirRNA sequences 50 nts upstream and 50 nts downstream (Appendix 1). The 146 pre-miRNA sequences were used as input for the ERPIN, ProMir and Mir-abela programs (see chapter 3). Random sequences of 240 nts from 200 known chicken protein coding genes and tRNAs, rRNA and snRNA genes with 50 nucleotides flanking each sites (Appendix 2) were selected as negative data inputs for ERPIN, ProMir and Mir-abela.

Sensitivity and Specificity

The fidelity of Mir-explore using chicken data was assessed by calculating the proportion of both positive and negative data that are correctly identified by any of the programs. Sensitivity was defined as the proportion of the known chicken pre-miRNA that is correctly identified by the algorithm. Specificity is the proportion of the negative data that is correctly identified by the algorithm.

4.4 Results

The sensitivity and specificity of mir-Explore together with ERPIN, ProMir-g and MiR-abela in detecting known chicken miRNA is shown in Table 4.1. The sensitivity of mir-Explore is 89% compared with 53% for ProMir-g and 57% for mir-Abela. The specificity of mir-Explore is 100%. For the mir classes available in ERPIN (25%), mir-

Explore predicted 46 out of the 52 known pre-miRNA, a sensitivity of 88%. However, the sensitivity of ERPIN and mir-Explore varied between chromosomes. On chromosome 1, ERPIN had 100% sensitivity whereas mir-Explore had 86 % sensitivity. However, on chromosome 4, ERPIN had 50% sensitivity whereas mir-Explore had 100% sensitivity.

Table 4.1 Sensitivity and Specificity of ProMir[1], ERPIN[2], MiR-abela[3] and MiR-Explore in predicting known chicken pre-microRNA.

	ProMir	ERPIN	Mir-abela	MiR-Explore
Sensitivity	53.42%	93%	57.53%	89.04%
Specificity	100%	100%	100%	100%

[1]http://cbit.snu.ac.kr/~ProMiR2
[2]http://tagc.univ-mrs.fr/erpin/ (Only 22 of 88 known miRNA classes were available in ERPIN)
[3]http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi

About 42% of the known chicken pre-MiRNA are located on the major chromosomes (1-4). Out of the 62 pre-miRNA located on the major chromosomes (1-4), mir-Explore predicted 90% (56 out of 62) whereas ProMir-g and mir-Abela predicted 38% and 51%, respectively. The summary of pre-miRNA predicted by mir-Explore on different chromosomes are presented on Table 4.2. The locations, accession number and identification of the pre-miRNAs and the negative data are presented in Appendices 3 and 4, respectively.

Table 4.2 Prediction of known chicken pre-miRNA using miR-Explore

Chromosome #	# of known miRNA	# of known miRNA Predicted by MiR-Explore	Percentage Predicted
1	24	20	83.33%
2	17	16	94.12%
3	9	8	88.89%
4	12	12	100.00%
5	1	1	100.00%
6	3	2	66.67%
7	3	3	100.00%
8	4	4	100.00%
9	2	2	100.00%
10	5	5	100.00%
11	2	2	100.00%
12	5	4	80.00%
13	3	2	66.67%
14	2	2	100.00%
15	3	2	66.67%
17	6	6	100.00%
18	1	1	100.00%
19	4	4	100.00%
20	3	3	100.00%
21	4	3	75.00%
23	5	5	100.00%
24	4	4	100.00%
26	6	5	83.33%
27	1	1	100.00%
28	3	1	33.33%
Z	9	7	77.78%
Un_Random	4	4	100.00%
10_random	1	1	100.00%
Not Validated	3		
No Information	1		

4.5 Discussion

The algorithm of miR-Explore is more sensitive in detecting chicken pre-miRNA than ProMir and miR-abela. This implies that known chicken pre-miRNA can be predicted with higher accuracy than the existing programs. All the programs evaluated to date did not produce on single false positive. The miR-Explore program is similar to

ERPIN. They both rely on profiling secondary structure and scoring, however, the scoring strategies are different. ERPIN score the linear sequence in the loop as well the helical secondary structure. Since the secondary structure is conserved than linear sequences, the advantage of scoring the loop in ERPIN cannot be fully evaluated since ERPIN has a limited number of mir classes. Nevertheless, the sensitivity of ERPIN is slightly higher (93% versus 89%) than miR-Explore in predicting chicken pre-miRNA from the limited ERPIN mir classes.

The amount of training data in secondary structure profiling can affect the sensitivity of a program that is dependent on the profile. In some chicken miRNA classes, the amount of miRNA used in the training data are limited. The more known miRNA that can be used, the more robust the consensus secondary structure is. The miR-Explore program relies on iteration to align a query input to the consensus secondary structure. It is possible to miss some gaps which will inevitably affect the scoring and as a consequence limit its accuracy. In addition, the computer time required to scan the genome in enormous and makes this program inefficient. A sophisticated genome scanning technique is needed instead of shifting the consensus structure. Finally, miR-Explore should be used to scan known human and mouse pre-miRNA and the chicken genome for novel pre-miRNA.

References:

Altschul SF, G.W., Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tools. Journal of Molecular Biology 215:403–410.

Bartel, D.P., and C.Z. Chen. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nature Reviews Genetics 5:396-400.

Bloomston, M., W.L. Frankel, F. Petrocca, S. Volinia, H. Alder, J.P. Hagan, C.G. Liu, D. Bhatt, C. Taccioli, and C.M. Croce. 2007. MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from normal pancreas and chronic pancreatitis. Jama-Journal of the American Medical Association 297:1901-1908.

Brown , J.R., and P. Sanseau. 2005. A computational view of microRNAs and their targets. Drug Discovery Today : Biosilico 10:595-601.

Chendrimada, T.P., K.J. Finn, X.J. Ji, D. Baillat, R.I. Gregory, S.A. Liebhaber, A.E. Pasquinelli, and R. Shiekhattar. 2007. MicroRNA silencing through RISC recruitment of eIF6. Nature 447:823-U1.

Darnell, D.K., S. Kaur, S. Stanislaw, J.K. Konieczka, T.A. Yatskievych, and P.B. Antin. 2006. MicroRNA expression during chick embryo development. Developmental Dynamics 235:3156-3165.

Debernardi, S., S. Skoulakis, G. Molloy, T. Chaplin, A. Dixon-McIver, and B.D. Young. 2007. MicroRNA miR-181a correlates with morphological sub-class of acute myeloid leukaemia and the expression of its target genes in global genome-wide analysis. Leukemia 21:912-916.

Gautheret, D., and A. Lambert. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. Journal of Molecular Biology 313:1003-1011.

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 Research 34:D140-D144.

Hurt, E.M., and W.L. Farrar. 2007. The STATus of miR-21 in apoptosis. Blood 110:1086-1087.

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 293:834-838.

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 15:2654-2659.

Lai, E.C., P. Tomancak, R.W. Williams, and G.M. Rubin. 2003. Computational identification of Drosophila microRNA genes. Genome Biology 4.

Lawrie, C.H. 2007. MicroRNAs and haematology: small molecules, big function. British Journal of Haematology 137:503-512.

Lee, Y., K. Jeon, J.T. Lee, S. Kim, and V.N. Kim. 2002. MicroRNA maturation: stepwise processing and subcellular localization. Embo Journal 21:4663-4670.

Lee, Y., C. Ahn, J.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 425:415-419.

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 & Development 17:991-1008.

Mathupala, S.P., S. Mittal, M. Guthikonda, and A.E. Sloan. 2007. MicroRNA and brain tumors: A cause and a cure? DNA and Cell Biology 26:301-310.

Miska, E.A., E. Alvarez-Saavedra, M. Townsend, A. Yoshii, N. Sestan, P. Rakic, M. Constantine-Paton, and H.R. Horvitz. 2004. Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biology 5.

Mott, J.L., S. Kobayashi, S.F. Bronk, and G.J. Gores. 2007. mir-29 regulates Mcl-1 protein expression and apoptosis. Oncogene 26:6133-6140.

Nam, J.W., J. Kim, S.K. Kim, and B.T. Zhang. 2006. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Research 34:W455-W458.

Nam, J.W., K.R. Shin, J.J. Han, Y. Lee, V.N. Kim, and B.T. Zhang. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research 33:3570-3581.

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 408:86-89.

Schwarz, D.S., Y. Tomari, and P.D. Zamore. 2004. The RNA-induced silencing complex is a Mg2+-dependent endonuclease. Current Biology 14:787-791.

Telernan, A.A., and S.M. Cohen. 2006. Drosophila lacking microRNA miR-278 are defective in energy homeostasis. Genes & Development 20:417-422.

Wang, Y.M., R. Medvid, C. Melton, R. Jaenisch, and R. Blelloch. 2007. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nature Genetics 39:380-385.

Xu, P.Z., S.Y. Vernooy, M. Guo, and B.A. Hay. 2003. The Drosophila MicroRNA mir-14 suppresses cell death and is required for normal fat metabolism. Current Biology 13:790-795.

Ye, G., Y. Zhang, B. Yang, and C. Peng. 2007. Regulation of trophoblast cell proliferation and apoptosis by microRNA-368. Placenta 28:A47-A47.

Zhang, B.H., X.P. Pan, Q.L. Wang, G.P. Cobb, and T.A. Anderson. 2006. Computational identification of microRNAs and their targets. Computational Biology and Chemistry 30:395-407.

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

Zhao, Y., and D. Srivastava. 2007. A developmental view of microRNA function. Trends in Biochemical Sciences 32:189-197.

CHAPTER 5

CONCLUSION

MicroRNAs (miRNAs) are endogenous 21 to 23 nucleotide molecules that are known to regulate about 30% of protein-coding genes through the RNA interference pathway. MicroRNAs are known to be involved in the regulation of development transition, promotion of cell proliferation and suppression of apoptosis, and tumor-genesis. Low level expression of some miRNA has limited their *in-situ* discovery. Such limitations could be ameliorated by *in-silico* methodologies. The efficacies of three major algorithms (ERPIN, ProMir and miR-abela) in detecting miRNA were evaluated using chicken pre-miRNA data. The sensitivity of ProMir and miR-abela were 53% and 57%, respectively. ERPIN has only 25% of the miRNA classes, and within the classes that were available the sensitivity was 93%. All the 3 algorithms had did not predict any false positives from 200 negative data. The efficiencies of the existing programs are low for chicken data and an efficient algorithm may be needed to predict novel chicken pre-miRNAs.

A new program miR-Explore has been developed to predict chicken pre-MiRNA. The program develops consensus secondary structures profile of miRNA classes and aligns query sequences and assigning a binary score to matches and mismatches. MiR-Explore shows 89% sensitivity and 100% specificity in identifying known chicken pre-

mirRNA. The efficiency of the computational time of mir-Explored can be improved to scan the chicken genome for novel pre-mirRNAs.

REFERENCES

Altschul SF, G. W., Miller W, Myers EW, Lipman DJ (1990). "Basic Local Alignment Search Tools." Journal of Molecular Biology **215**(3): 403–410.

Ambros, V. (1989). "A Hierarchy of Regulatory Genes Controls a Larva-to-Adult Developmental Switch in C-Elegans." Cell **57**(1): 49-57.

Ambros, V. and H. R. Horvitz (1984). "Heterochronic Mutants of the Nematode Caenorhabditis-Elegans." Science **226**(4673): 409-416.

Amort, M., B. Wotzel, et al. (2007). "An intact ribose moiety at A2602 of 23S rRNA is key to trigger peptidyl-tRNA hydrolysis during translation termination." Nucleic Acids Research **35**(15): 5130-5140.

Armougom, F., S. Moretti, et al. (2006). "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-coffee." Nucleic Acids Research **34**: W604-W608.

Aronova, A., D. Bacikova, et al. (2007). "Functional interactions between Prp8, Prp18, Slu7, and U5 snRNA during the second step of pre-mRNA splicing." Rna-a Publication of the Rna Society **13**(9): 1437-1444.

Bagga, S., J. Bracht, et al. (2005). "Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation." Cell **122**(4): 553-563.

Bartel, D. P. (2004). "MicroRNAs: Genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Bartel, D. P. and C. Z. Chen (2004). "Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs." Nature Reviews Genetics **5**(5): 396-400.

Bellman, R. (1955). "Dynamic Programming." Journal of the Operations Research Society of America **3**(3): 352-352.

Bentwich, I., A. Avniel, et al. (2005). "Identification of hundreds of conserved and nonconserved human microRNAs." Nature Genetics **37**(7): 766-770.

Berezikov, E., V. Guryev, et al. (2005). "Phylogenetic shadowing and computational identification of human microRNA genes." Cell **120**(1): 21-24.

Bertram, G., S. Innes, et al. (2001). "Endless possibilities: translation termination and stop codon recognition." Microbiology-Uk **147**: 255-269.

Blackshields G, W. I., Larkin M, Higgins DG (2006). "Analysis and comparison of benchmarks for multiple sequence alignment." In Silico Biology **6**: 321-339.

Bloomston, M., W. L. Frankel, et al. (2007). "MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from normal pancreas and chronic pancreatitis." Jama-Journal of the American Medical Association **297**(17): 1901-1908.

Boffelli, D., J. McAuliffe, et al. (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-1394.

Bonnet, E., J. Wuyts, et al. (2004). "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences." Bioinformatics **20**(17): 2911-2917.

Borukhov, S. and E. Nudler (2003). "RNA polymerase holoenzyme: structure, function and biological implications." Current Opinion in Microbiology **6**(2): 93-100.

Brennecke, J., D. R. Hipfner, et al. (2003). "bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila." Cell **113**(1): 25-36.

Brown , J. R. and P. Sanseau (2005). "A computational view of microRNAs and their targets." Drug Discovery Today : Biosilico **10**(8): 595-601.

Busch, H., R. Reddy, et al. (1982). "Snrnas, Snrnps, and Rna Processing." Annual Review of Biochemistry **51**: 617-654.

Calin, G. A., C. D. Dumitru, et al. (2002). "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia." Proceedings of the National Academy of Sciences of the United States of America **99**(24): 15524-15529.

Calin, G. A., C. Sevignani, et al. (2004). "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers." Proceedings of the National Academy of Sciences of the United States of America **101**(9): 2999-3004.

Carthew, R. W. (2001). "Gene silencing by double-stranded RNA." Current Opinion in Cell Biology **13**(2): 244-248.

Chen, C. Z., L. Li, et al. (2004). "MicroRNAs modulate hematopoietic lineage differentiation." Science **303**(5654): 83-86.

Chen, J. H., S. Y. Le, et al. (2000). "Prediction of common secondary structures of RNAs: a genetic algorithm approach." Nucleic Acids Research **28**(4): 991-999.

Chendrimada, T. P., K. J. Finn, et al. (2007). "MicroRNA silencing through RISC recruitment of eIF6." Nature **447**(7146): 823-U1.

Cimmino, A., G. A. Calin, et al. (2005). "miR-15 and miR-16 induce apoptosis by targeting BCL2." Proceedings of the National Academy of Sciences of the United States of America **102**(39): 13944-13949.

Crick, F. (1970). "Central Dogma of Molecular Biology." Nature **227**(5258): 561-&.

Cullen, B. R. (2006). "Viruses and microRNAs." Nature Genetics **38**: S25-S30.

Darnell, D. K., S. Kaur, et al. (2006). "MicroRNA expression during chick embryo development." Developmental Dynamics **235**(11): 3156-3165.

Debernardi, S., S. Skoulakis, et al. (2007). "MicroRNA miR-181a correlates with morphological sub-class of acute myeloid leukaemia and the expression of its target genes in global genome-wide analysis." Leukemia **21**(5): 912-916.

Do, C. B., M. S. P. Mahabhashyam, et al. (2005). "ProbCons: Probabilistic consistency-based multiple sequence alignment." Genome Research **15**(2): 330-340.

Donker, R. B., J. F. Mouillet, et al. (2007). "The expression of Argonaute2 and related microRNA biogenesis proteins in normal and hypoxic trophoblasts." Molecular Human Reproduction **13**(4): 273-279.

Du, T. T. and P. D. Zamore (2005). "microPrimer: The biogenesis and function of microRNA." Development **132**(21): 4645-4652.

Dumas, J. P. and J. Ninio (1982). "Efficient Algorithms for Folding and Comparing Nucleic-Acid Sequences." Nucleic Acids Research **10**(1): 197-206.

Eddy, S. R. (2001). "Non-coding RNA genes and the modern RNA world." Nature Reviews Genetics **2**(12): 919-929.

Edgar, R. C. (2004). "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." Bmc Bioinformatics **5**: 1-19.

Edgar, R. C. and S. Batzoglou (2006). "Multiple sequence alignment." Current Opinion in Structural Biology **16**(3): 368-373.

Elbashir, S. M., W. Lendeckel, et al. (2001). "RNA interference is mediated by 21-and 22-nucleotide RNAs." Genes & Development **15**(2): 188-200.

Eliceiri, G. L. (1999). "Small nucleolar RNAs." Cellular and Molecular Life Sciences **56**(1-2): 22-31.

Erdei, S., I. Boros, et al. (1983). "A Novel Type of Bacterial Transcription Unit, Specifying Messenger-Rna, Ribosomal-Rna, and Transfer-Rna." Molecular & General Genetics **191**(1): 162-164.

Feng, D. F. and R. F. Doolittle (1987). "Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees." Journal of Molecular Evolution **25**(4): 351-360.

Freier, S. M., R. Kierzek, et al. (1986). "Improved Free-Energy Parameters for Predictions of Rna Duplex Stability." Proceedings of the National Academy of Sciences of the United States of America **83**(24): 9373-9377.

Gautheret, D. and A. Lambert (2001). "Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles." Journal of Molecular Biology **313**(5): 1003-1011.

Gilbert, W. (1978). "Why Genes in Pieces." Nature **271**(5645): 501-501.

Gregory, R. I. and R. Shiekhattar (2005). "MicroRNA biogenesis and cancer." Cancer Research **65**(9): 3509-3512.

Griffiths-Jones, S. (2004). "The microRNA Registry." Nucleic Acids Research **32**: D109-D111.

Griffiths-Jones, S., R. J. Grocock, et al. (2006). "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic Acids Research **34**: D140-D144.

Griffiths-Jones, S., S. Moxon, et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Research **33**: D121-D124.

Hariharan, M., V. Scaria, et al. (2005). "Targets for human encoded mircoRNAs in HIV genes." Biochemical and Biophysical Research Communications **337**(4): 1214-1218.

Harvey Lodish, A. B., Lawrence S. Zipursky, Paul Matsudaira, David Baltimore, James Darnell (2000). Molecular Cell Biology. New York, W. H. FREEMAN.

He, L. and G. J. Hannon (2004). "Microrna: Small RNAs with a big role in gene regulation." Nature Reviews Genetics **5**(7): 522-531.

He, L., J. M. Thomson, et al. (2005). "A microRNA polycistron as a potential human oncogene." Nature **435**(7043): 828-833.

Hipfner, D. R., K. Weigmann, et al. (2002). "The bantam gene regulates Drosophila growth." Genetics **161**(4): 1527-1537.

Hofacker, I. L., W. Fontana, et al. (1994). "Fast Folding and Comparison of Rna Secondary Structures." Monatshefte Fur Chemie **125**(2): 167-188.

Horvitz, H. R. and J. E. Sulston (1980). "Isolation and Genetic-Characterization of Cell-Lineage Mutants of the Nematode Caenorhabditis-Elegans." Genetics **96**(2): 435-454.

Humphreys, D. T., B. J. Westman, et al. (2005). "MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function." Proceedings of the National Academy of Sciences of the United States of America **102**(47): 16961-16966.

Hurt, E. M. and W. L. Farrar (2007). "The STATus of miR-21 in apoptosis." Blood **110**(4): 1086-1087.

Hutvagner, G., J. McLachlan, et al. (2001). "A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA." Science **293**(5531): 834-838.

Hutvagner, G., M. J. Simard, et al. (2004). "Sequence-specific inhibition of small RNA function." Plos Biology **2**(4): 465-475.

Hutvagner, G. and P. D. Zamore (2002). "A microRNA in a multiple-turnover RNAi enzyme complex." Science **297**(5589): 2056-2060.

Jovanovic, M. and M. O. Hengartner (2006). "miRNAs and apoptosis: RNAs to die for." Oncogene **25**(46): 6176-6187.

Ketting, R. F., S. E. J. Fischer, et al. (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, et al. (2003). "Functional siRNAs and miRNAs exhibit strand bias (vol 115, pg 209, 2003)." Cell **115**(4): 505-505.

Kim, J., A. Krichevsky, et al. (2004). "Identification of many microRNAs that copurify with polyribosomes in mammalian neurons." Proceedings of the National Academy of Sciences of the United States of America **101**(1): 360-365.

Kim, V. N. (2005). "Small RNAs: Classification, biogenesis, and function." Molecules and Cells **19**(1): 1-15.

Kozak, M. (1986). "Point Mutations Define a Sequence Flanking the Aug Initiator Codon That Modulates Translation by Eukaryotic Ribosomes." Cell **44**(2): 283-292.

Krichevsky, A. M., K. S. King, et al. (2003). "A microRNA array reveals extensive regulation of microRNAs during brain development." Rna-a Publication of the Rna Society **9**(10): 1274-1281.

Kumar, S. and A. Rzhetsky (1996). "Evolutionary relationships of eukaryotic kingdoms." Journal of Molecular Evolution **42**(2): 183-193.

Lai, E. C., P. Tomancak, et al. (2003). "Computational identification of Drosophila microRNA genes." Genome Biology **4**(7).

Lambert, A., J. F. Fontaine, et al. (2004). "The ERPIN server: an interface to profile-based RNA motif identification." Nucleic Acids Research **32**: W160-W165.

Landthaler, M., A. Yalcin, et al. (2004). "The human DiGeorge syndrome critical region gene 8 and its D-melanogaster homolog are required for miRNA biogenesis." Current Biology **14**(23): 2162-2167.

Lau, N. C., L. P. Lim, et al. (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.

Lawrie, C. H. (2007). "MicroRNAs and haematology: small molecules, big function." British Journal of Haematology **137**(6): 503-512.

Lee, J., Z. H. Li, et al. (2007). "Regulatory circuit of human microRNA biogenesis." Plos Computational Biology **3**(4): 721-732.

Lee, R. C. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-864.

Lee, R. C., R. L. Feinbaum, et al. (1993). "The C-Elegans Heterochronic Gene Lin-4 Encodes Small Rnas with Antisense Complementarity to Lin-14." Cell **75**(5): 843-854.

Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.

Lee, Y., K. Jeon, et al. (2002). "MicroRNA maturation: stepwise processing and subcellular localization." Embo Journal **21**(17): 4663-4670.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." Embo Journal **23**(20): 4051-4060.

Legendre, M., A. Lambert, et al. (2005). "Profile-based detection of microRNA precursors in animal genomes." Bioinformatics **21**(7): 841-845.

Lewis, J. D. and E. Izaurralde (1997). "The role of the cap structure in RNA processing and nuclear export." European Journal of Biochemistry **247**(2): 461-469.

Lim, L. P., N. C. Lau, et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." Nature **433**(7027): 769-773.

Lim, L. P., N. C. Lau, et al. (2003). "The microRNAs of Caenorhabditis elegans." Genes & Development **17**(8): 991-1008.

Lund, E., S. Guttinger, et al. (2004). "Nuclear export of microRNA precursors." Science **303**(5654): 95-98.

Mathews, D. H., J. Sabina, et al. (1999). "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." Journal of Molecular Biology **288**(5): 911-940.

Mathews, D. H. and D. H. Turner (2002). "Dynalign: An algorithm for finding the secondary structure common to two RNA sequences." Journal of Molecular Biology **317**(2): 191-203.

Mathupala, S. P., S. Mittal, et al. (2007). "MicroRNA and brain tumors: A cause and a cure?" DNA and Cell Biology **26**(5): 301-310.

Mattick, J. S. and I. V. Makunin (2006). "Non-coding RNA." Human Molecular Genetics **15**: R17-R29.

Metzler, M., M. Wilda, et al. (2004). "High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma." Genes Chromosomes & Cancer **39**(2): 167-169.

Michael, M. Z., S. M. O'Connor, et al. (2003). "Reduced accumulation of specific microRNAs in colorectal neoplasia." Molecular Cancer Research **1**(12): 882-891.

Miska, E. A., E. Alvarez-Saavedra, et al. (2004). "Microarray analysis of microRNA expression in the developing mammalian brain." Genome Biology **5**(9).

Mott, J. L., S. Kobayashi, et al. (2007). "mir-29 regulates Mcl-1 protein expression and apoptosis." Oncogene **26**(42): 6133-6140.

Nam, J. W., J. Kim, et al. (2006). "ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs." Nucleic Acids Research **34**(Web Server): W455-W458.

Nam, J. W., J. Kim, et al. (2006). "ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs." Nucleic Acids Research **34**: W455-W458.

Nam, J. W., K. R. Shin, et al. (2005). "Human microRNA prediction through a probabilistic co-learning model of sequence and structure." Nucleic Acids Research **33**(11): 3570-3581.

Nelson, P., M. Kiriakidou, et al. (2003). "The microRNA world: small is mighty." Trends in Biochemical Sciences **28**(10): 534-540.

Nelson, P. T., A. G. Hatzigeorgiou, et al. (2004). "miRNP : mRNA association in polyribosomes in a human neuronal cell line." Rna-a Publication of the Rna Society **10**(3): 387-394.

Nilsson, L. and O. Nygard (1984). "Affinity Labeling of the Eukaryotic Elongation-Factor Ef-2 with the Guanosine Nucleotide Analog 5'-Para-Fluorosulfonylbenzoylguanosine." Biochimica Et Biophysica Acta **782**(1): 49-54.

Notredame, C. (2002). "Recent progress in multiple sequence alignment: a survey." Pharmacogenomics **3**(1): 131-144.

Notredame, C. (2007). "Recent evolutions of multiple sequence alignment algorithms." Plos Computational Biology **3**(8): 1405-1408.

Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." Journal of Molecular Biology **302**(1): 205-217.

Olivas, W. M., D. Muhlrad, et al. (1997). "Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs." Nucleic Acids Research **25**(22): 4619-4625.

Olsen, P. H. and V. Ambros (1999). "The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation." Developmental Biology **216**(2): 671-680.

Ouellet, D. L., M. P. Perron, et al. (2006). "MicroRNAs in gene regulation: When the smallest governs it all." Journal of Biomedicine and Biotechnology: 168.

Pasquinelli, A. E., B. J. Reinhart, et al. (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA." Nature **408**(6808): 86-89.

Pei, J. M. and N. V. Grishin (2006). "MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information." Nucleic Acids Research **34**(16): 4364-4374.

Pei, J. M., R. Sadreyev, et al. (2003). "PCMA: fast and accurate multiple sequence alignment based on profile consistency." Bioinformatics **19**(3): 427-428.

Pfeffer, S., M. Zavolan, et al. (2004). "Identification of virus-encoded microRNAs." Science **304**(5671): 734-736.

Pillai, R. S. (2005). "MicroRNA function: Multiple mechanisms for a tiny RNA?" Rna-a Publication of the Rna Society **11**(12): 1753-1761.

Poy, M. N., L. Eliasson, et al. (2004). "A pancreatic islet-specific microRNA regulates insulin secretion." Nature **432**(7014): 226-230.

Reinhart, B. J. and D. P. Bartel (2002). "Small RNAs correspond to centromere heterochromatic repeats." Science **297**(5588): 1831-1831.

Reinhart, B. J., F. J. Slack, et al. (2000). "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans." Nature **403**(6772): 901-906.

Ritchie, W., M. Legendre, et al. (2007). "RNA stem-loops: To be or not to be cleaved by RNAse III." Rna-a Publication of the Rna Society **13**(4): 457-462.

Rodriguez, A., S. Griffiths-Jones, et al. (2004). "Identification of mammalian microRNA host genes and transcription units." Genome Research **14**(10A): 1902-1910.

Russell, P. J. (2006). iGenetics A Molecular Approach, Pearson Education Inc.

Ryther, R. C. C., A. S. Flynt, et al. (2005). "siRNA therapeutics: Big potential from small RNAs." Gene Therapy **12**(1): 5-11.

Sachidanandam, R. (2005). "RNAi as a bioinformatics consumer." Briefings in Bioinformatics **6**(2): 146-162.

Samols, M. A., J. H. Hu, et al. (2005). "Cloning and identification of a MicroRNA cluster within the latency-associated region of Kaposi's sarcoma-associated herpesvirus." Journal of Virology **79**(14): 9301-9305.

Schwarz, D. S., G. Hutvagner, et al. (2003). "Asymmetry in the assembly of the RNAi enzyme complex." Cell **115**(2): 199-208.

Schwarz, D. S., Y. Tomari, et al. (2004). "The RNA-induced silencing complex is a Mg2+-dependent endonuclease." Current Biology **14**(9): 787-791.

Seggerson, K., L. J. Tang, et al. (2002). "Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation." Developmental Biology **243**(2): 215-225.

Seitz, H., N. Youngson, et al. (2003). "Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene." Nature Genetics **34**(3): 261-262.

Sewer, A., N. Paul, et al. (2005). "Identification of clustered microRNAs using an ab initio prediction method." Bmc Bioinformatics **6**.

Sharp, P. A. (2001). "RNA interference - 2001." Genes & Development **15**(5): 485-490.

Shine, J. and L. Dalgarno (1975). "Terminal-Sequence Analysis of Bacterial Ribosomal-Rna Correlation between 3'-Terminal Polypyrimidine Sequence of 16-S Rna and Translational Specificity of Ribosome." European Journal of Biochemistry **57**(1): 221-230.

Silverman, A. (2003). A Critical Review of Computational Methods for RNA

Secondary Structure Prediction. Biochem218 Standford University.

Stevenson, M. (2004). "Therapeutic potential of RNA interference." New England Journal of Medicine **351**(17): 1772-1777.

Takamizawa, J., H. Konishi, et al. (2004). "Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival." Cancer Research **64**(11): 3753-3756.

Taylor, W. R. (1986). "Identification of Protein-Sequence Homology by Consensus Template Alignment." Journal of Molecular Biology **188**(2): 233-258.

Telernan, A. A. and S. M. Cohen (2006). "Drosophila lacking microRNA miR-278 are defective in energy homeostasis." Genes & Development **20**(4): 417-422.

Thomassen, G. O. S., O. Rosok, et al. (2006). "Computational prediction of microRNAs encoded in viral and other genomes." Journal of Biomedicine and Biotechnology.

Thompson, J. D., D. G. Higgins, et al. (1994). "Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research **22**(22): 4673-4680.

Thompson, J. D., F. Plewniak, et al. (2000). "DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches." Nucleic Acids Research **28**(15): 2919-2926.

Vance, V. and H. Vaucheret (2001). "RNA silencing in plants - Defense and counterdefense." Science **292**(5525): 2277-2280.

Wallace, I. M., O. O'Sullivan, et al. (2006). "M-Coffee: combining multiple sequence alignment methods with T-Coffee." Nucleic Acids Research **34**(6): 1692-1699.

Walter, A. E., D. H. Turner, et al. (1994). "Coaxial Stacking of Helixes Enhances Binding of Oligoribonucleotides and Improves Predictions of Rna Folding." Proceedings of the National Academy of Sciences of the United States of America **91**(20): 9218-9222.

Wang, J. F., H. Zhou, et al. (2004). "Identification of 20 microRNAs from Oryza sativa." Nucleic Acids Research **32**(5): 1688-1695.

Wang, L. a. J., T. (1994). "On the complexity of multiple sequence alignment." Computational Biology(1): 337-348.

Wang, X. W., J. Zhang, et al. (2005). "MicroRNA identification based on sequence and structure alignment." Bioinformatics **21**(18): 3610-3614.

Wang, Y. M., R. Medvid, et al. (2007). "DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal." Nature Genetics **39**(3): 380-385.

Watson, J. D. and F. H. C. Crick (1953). "The Structure of DNA." Cold Spring Harbor Symposia on Quantitative Biology **18**: 123-131.

Wightman, B., I. Ha, et al. (1993). "Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern-Formation in C-Elegans." Cell **75**(5): 855-862.

Xia, T. B., J. SantaLucia, et al. (1998). "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs." Biochemistry **37**(42): 14719-14735.

Xu, P. Z., S. Y. Vernooy, et al. (2003). "The Drosophila MicroRNA mir-14 suppresses cell death and is required for normal fat metabolism." Current Biology **13**(9): 790-795.

Ye, G., Y. Zhang, et al. (2007). "Regulation of trophoblast cell proliferation and apoptosis by microRNA-368." Placenta **28**(8-9): A47-A47.

Zeng, Y. and B. R. Cullen (2004). "Structural requirements for pre-microRNA binding and nuclear export by Exportin 5." Nucleic Acids Research **32**(16): 4776-4785.

Zeng, Y. and B. R. Cullen (2005). "Efficient processing of primary microRNA hairpins by drosha requires flanking nonstructured RNA sequences." Journal of Biological Chemistry **280**(30): 27595-27603.

Zeng, Y., R. Yi, et al. (2003). "MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms." Proceedings of the National Academy of Sciences of the United States of America **100**(17): 9779-9784.

Zeng, Y., R. Yi, et al. (2005). "Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha." Embo Journal **24**(1): 138-148.

Zhang, B. H., X. P. Pan, et al. (2006). "Computational identification of microRNAs and their targets." Computational Biology and Chemistry **30**(6): 395-407.

Zhang, H. D., F. A. Kolb, et al. (2004). "Single processing center models for human dicer and bacterial RNase III." Cell **118**(1): 57-68.

Zhao, Y. and D. Srivastava (2007). "A developmental view of microRNA function." Trends in Biochemical Sciences **32**(4): 189-197.

Zieve, G. W. (1981). "2 Groups of Small Stable Rnas." Cell **25**(2): 296-297.

Zuker, M. (1989). "On Finding All Suboptimal Foldings of an Rna Molecule." Science **244**(4900): 48-52.

Zuker, M. (1994). "Prediction of RNA secondary structure by energy minimization." Methods in Molecular Biology **25**: 267-294.

APPENDICES

Appendix 1.
Detail result of using ProMir, ERPIN and miR-abela to identify known chicken pre-miRNAs (sorted based on the Accession ID in each chromosome)

Chromosome #	Position	Accession	ID	ProMir II	ERPIN	miR-abela
1	3,236,279 - 3,236,467	MI0001166	gga-mir-29a	Detected	Detected	Detected
	3,235,262 - 3,235,442	MI0001167	gga-mir-29b-1	Detected	Detected	Detected
	34,895,637 - 34,895,820	MI0001168	gga-let-7i	Detected	Detected	Detected
	48,192,609 - 48,192,808	MI0001169	gga-mir-135a-2	Detected	N/A	Detected
	51,372,232 - 51,372,400	MI0001170	gga-mir-33	Not Detected	N/A	Not Detected
	73,421,222 - 73,421,397	MI0001171	gga-let-7a-3	Detected	Detected	Detected
	73,422,051 - 73,422,235	MI0001172	gga-let-7b	Not Detected	Detected	Not Detected
	102,424,283 - 102,424,463	MI0001173	gga-mir-99a	Detected	Detected	Detected
	102,425,036 - 102,425,219	MI0001174	gga-let-7c	Detected	Detected	Detected
	102,457,597 - 102,457,786	MI0001175	gga-mir-125b	Detected	Detected	Detected
	105,930,163 - 105,930,325	MI0001176	gga-mir-155	Detected	N/A	Detected
	114,215,977 - 114,216,174	MI0001177	gga-mir-222	Not Detected	N/A	Not Detected
	114,218,372 - 114,218,569	MI0001177	gga-mir-222	Detected	N/A	Not Detected
	114,218,876 - 114,219,074	MI0001178	gga-mir-221	Detected	N/A	Detected
	152,248,020 - 152,248,197	MI0001179	gga-mir-92	Not Detected	Detected	Not Detected
	152,248,133 - 152,248,319	MI0001180	gga-mir-19b	Detected	N/A	Not Detected
	152,248,256 - 152,248,453	MI0001181	gga-mir-20a	Not Detected	Detected	Detected
	152,248,442 - 152,248,572	MI0001182	gga-mir-19a	Not Detected	N/A	Not Detected
	152,248,576 - 152,248,768	MI0001183	gga-mir-18a	Not Detected	Detected	Not Detected
	152,248,731 - 152,248,915	MI0001184	gga-mir-17	Not Detected	Detected	Detected
	173,700,301 - 173,700,484	MI0001185	gga-mir-16-1	Not Detected	Detected	Not Detected
	173,700,443 - 173,700,625	MI0001186	gga-mir-15a	Not Detected	Detected	Not Detected

Appendix 1 continued

	1,147,470 - 1,147,667	MI0001238	gga-mir-205b	Detected	N/A	Detected
	59,948,651 - 59,948,843	MI0003709	gga-mir-490	Not Detected	N/A	Not Detected
2	4,467,466 - 4,467,642	MI0001187	gga-mir-26a	Not Detected	N/A	Detected
	8,765,637 - 8,765,823	MI0001188	gga-mir-153	Not Detected	N/A	Detected
	32,053,493 - 32,053,660	MI0001189	gga-mir-148a	Not Detected	N/A	Not Detected
	32,586,099 - 32,586,292	MI0001190	gga-mir-196-2	Not Detected	N/A	Not Detected
	40,745,098 - 40,745,293	MI0001191	gga-mir-138-1	Not Detected	N/A	Not Detected
	45,549,126 - 45,549,309	MI0001192	gga-mir-128-2	Not Detected	N/A	Detected
	85,892,420 - 85,892,605	MI0001193	gga-mir-187	Detected	N/A	Detected
	86,506,401 - 86,506,570	MI0001194	gga-mir-32	Not Detected	N/A	Not Detected
	105,670,307 - 105,670,493	MI0001195	gga-mir-133a-1	Not Detected	Detected	Detected
	105,673,433 - 105,673,617	MI0001196	gga-mir-1a-2	Not Detected	Detected	Not Detected
	118,524,107 - 118,524,302	MI0001197	gga-mir-124a	Not Detected	Detected	Detected
	148,337,213 - 148,337,376	MI0001198	gga-mir-30d	Not Detected	Detected	Not Detected
	148,331,548 - 148,331,734	MI0001199	gga-mir-30b	Not Detected	Detected	Not Detected
	23,068,827 - 23,069,010	MI0003708	gga-mir-489	Not Detected	N/A	Not Detected
	21,671,911 - 21,672,079	MI0003713	gga-mir-466	Not Detected	N/A	Not Detected
	3,583,640 - 3,583,829	MI0004998	gga-mir-460	Not Detected	N/A	Not Detected
	78,310,794 - 78,310,975	MI0004999	gga-mir-757-1	Not Detected	N/A	Not Detected
3	282,757 - 282,962	MI0001200	gga-mir-216	Not Detected	N/A	Not Detected
	280,089 - 280,295	MI0001201	gga-mir-217	Detected	N/A	Not Detected
	19,924,437 - 19,924,611	MI0001202	gga-mir-194	Not Detected	N/A	Detected
	19,924,743 - 19,924,947	MI0001203	gga-mir-215	Not Detected	N/A	Detected
	85,102,189 - 85,102,360	MI0001204	gga-mir-30a	Detected	Detected	Detected
	85,126,803 - 85,126,974	MI0001205	gga-mir-30c-2	Not Detected	Detected	Not Detected
	110,384,883 - 110,385,066	MI0001206	gga-mir-133b	Not Detected	Detected	Detected
	110,390,389 - 110,390,564	MI0001207	gga-mir-206	Not Detected	Detected	Not Detected
	32,679,660 - 32,679,871	MI0004997	gga-mir-456	Not Detected	N/A	Not Detected
4	232,899 - 233,098	MI0001208	gga-mir-223	Detected	Detected	Detected
	3,970,178 - 3,970,361	MI0001209	gga-mir-18b	Not Detected	Detected	Detected
	3,970,309 - 3,970,489	MI0001210	gga-mir-106	Detected	Not Detected	Detected

Appendix 1 continued

	58,651,829 - 58,651,995	MI0001211	gga-mir-302a	Detected	Detected	Detected
	77,774,648 - 77,774,856	MI0001212	gga-mir-218-1	Detected	N/A	Detected
	91,906,839 - 91,907,021	MI0001213	gga-mir-103-2	Not Detected	N/A	Not Detected
	3,969,997 - 3,970,181	MI0001517	gga-mir-20b	Not Detected	Detected	Not Detected
	58,651,264 - 58,651,435	MI0003700	gga-mir-302b	Detected	Not Detected	Detected
	58,651,526 - 58,651,690	MI0003701	gga-mir-302c	Detected	Not Detected	Detected
	58,652,164 - 58,652,332	MI0003702	gga-mir-302d	Detected	Not Detected	Detected
	65,844,645 - 65,844,817	MI0003706	gga-mir-383	Not Detected	N/A	Not Detected
	58,652,310 - 58,652,472	MI0003712	gga-mir-367	Detected	N/A	Detected
5	53,206,764 - 53,206,961	MI0001214	gga-mir-203	Not Detected	Detected	Detected
6	20,487,914 - 20,488,094	MI0001215	gga-mir-107	Not Detected	N/A	Not Detected
	24,570,010 - 24,570,214	MI0003695	gga-mir-146b	Detected	N/A	Detected
	22,813,018 - 22,813,206	MI0003699	gga-mir-202	Detected	N/A	Detected
7	17,388,998 - 17,389,207	MI0001216	gga-mir-10b	Not Detected	Detected	Not Detected
	32,228,100 - 32,228,281	MI0001217	gga-mir-128-1	Not Detected	N/A	Detected
	23,901,074 - 23,901,238	MI0003705	gga-mir-375	Not Detected	N/A	Detected
8	2,001,511 - 2,001,714	MI0001218	gga-mir-181a-1	Detected	N/A	Detected
	2,001,700 - 2,001,888	MI0001219	gga-mir-181b-1	Detected	N/A	Not Detected
	4,732,723 - 4,732,930	MI0001220	gga-mir-199-2	Not Detected	N/A	Detected
	13,210,143 - 13,210,338	MI0001221	gga-mir-137	Not Detected	N/A	Detected
9	23,742,741 - 23,742,934	MI0001222	gga-mir-16-2	Detected	Detected	Detected
	23,742,916 - 23,743,106	MI0001223	gga-mir-15b	Detected	Detected	Not Detected
10	5,209,674 - 5,209,858	MI0001224	gga-mir-190	Not Detected	N/A	Detected
	6,651,224 - 6,651,424	MI0001225	gga-mir-204-2	Detected	N/A	Detected
	14,823,475 - 14,823,673	MI0001226	gga-mir-7-2	Not Detected	N/A	Not Detected
	22,146,195 - 22,146,368	MI0001227	gga-mir-184	Detected	N/A	Detected
	12,334,872 - 12,335,041	MI0003697	gga-mir-147-2	Not Detected	N/A	Detected
11	2,023,904 - 2,024,086	MI0001228	gga-mir-138-2	Not Detected	N/A	Not Detected

Appendix 1 continued

	21,030,591 - 21,030,785	MI0001229	gga-mir-140	Detected	Detected	Detected
12	2,809,028 - 2,809,210	MI0001230	gga-let-7g	Detected	Detected	Detected
	2,830,692 - 2,830,879	MI0001231	gga-mir-135a-1	Not Detected	N/A	Not Detected
	6,301,402 - 6,301,604	MI0001232	gga-let-7d	Detected	Detected	Detected
	6,302,447 - 6,302,633	MI0001233	gga-let-7f	Not Detected	Detected	Not Detected
	6,302,861 - 6,303,050	MI0001234	gga-let-7a-1	Detected	Detected	Not Detected
13	7,555,543 - 7,555,741	MI0001235	gga-mir-146a	Detected	N/A	Detected
	4,449,192 - 4,449,369	MI0001236	gga-mir-103-1	Detected	N/A	Detected
	4,322,810 - 4,323,004	MI0001237	gga-mir-218-2	Not Detected	N/A	Detected
14	759,403 - 759,585	MI0003698	gga-mir-193	Not Detected	N/A	Not Detected
	764,221 - 764,405	MI0003703	gga-mir-365-1	Detected	N/A	Not Detected
15	398,670 - 398,846	MI0001239	gga-mir-130b	Not Detected	N/A	Detected
	406,263 - 406,455	MI0001240	gga-mir-301	Detected	N/A	Not Detected
	408,349 - 408,531	MI0001241	gga-mir-130a	Detected	N/A	Not Detected
17	10,220,087 - 10,220,271	MI0001242	gga-mir-181b-2	Detected	N/A	Detected
	10,218,447 - 10,218,637	MI0001243	gga-mir-181a-2	Not Detected	N/A	Detected
	8,431,692 - 8,431,875	MI0001244	gga-mir-126	Detected	N/A	Not Detected
	5,667,100 - 5,667,293	MI0001245	gga-mir-199-1	Detected	N/A	Detected
	5,577,767 - 5,577,951	MI0001246	gga-mir-219	Detected	N/A	Detected
	5,339,651 - 5,339,836	MI0003707	gga-mir-455	Detected	N/A	Detected
18	6,437,246 - 6,437,441	MI0003704	gga-mir-365-2	Detected	N/A	Detected
19	496,933 - 497,120	MI0001281	gga-mir-142	Detected	N/A	Detected
	7,322,022 - 7,322,218	MI0004994	gga-mir-21	Detected	N/A	Detected
	5,823,918 - 5,824,086	MI0004995	gga-mir-451	Detected	N/A	Detected
	5,824,073 - 5,824,257	MI0004996	gga-mir-144	Detected	N/A	Not Detected
20	8,107,781 - 8,107,951	MI0001247	gga-mir-1a-1	Detected	Detected	Detected

Appendix 1 continued

	8,119,004 - 8,119,199	MI0001248	gga-mir-133a-2	Detected	Detected	Detected
	2,599,284 - 2,599,474	MI0003710	gga-mir-499	Detected	N/A	Detected
21	2,583,267 - 2,583,453	MI0001249	gga-mir-200a	Detected	N/A	Detected
	2,585,592 - 2,585,776	MI0001250	gga-mir-200b	Detected	N/A	Not Detected
	3,251,464 - 3,251,672	MI0001251	gga-mir-34a	Not Detected	Detected	Not Detected
	2,580,762 - 2,580,945	MI0003714	gga-mir-429	Not Detected	N/A	Not Detected
23	2,510,281 - 2,510,473	MI0001252	gga-mir-124b-1	Detected	Detected	Detected
	4,663,862 - 4,664,025	MI0001254	gga-mir-1b	Detected	Detected	Not Detected
	4,664,001 - 4,664,179	MI0001255	gga-mir-133c	Not Detected	Detected	Not Detected
	5,248,364 - 5,248,559	MI0001256	gga-mir-30e	Detected	Detected	Detected
	5,249,587 - 5,249,775	MI0001257	gga-mir-30c-1	Not Detected	Detected	Detected
24	3,372,844 - 3,373,023	MI0001258	gga-mir-100	Detected	Detected	Detected
	3,380,943 - 3,381,114	MI0001259	gga-let-7a-2	Detected	Detected	Detected
	5,684,850 - 5,685,033	MI0001260	gga-mir-34b	Detected	Detected	Detected
	5,685,587 - 5,685,760	MI0001261	gga-mir-34c	Detected	Detected	Detected
26	1,442,647 - 1,442,829	MI0001262	gga-let-7j	Not Detected	Detected	Not Detected
	1,442,847 - 1,443,029	MI0001263	gga-let-7k	Detected	Detected	Not Detected
	1,925,892 - 1,926,087	MI0001264	gga-mir-135a-3	Not Detected	N/A	Not Detected
	2,511,608 - 2,511,796	MI0001265	gga-mir-29c	Not Detected	Detected	Not Detected
	2,512,519 - 2,512,698	MI0001266	gga-mir-29b-2	Not Detected	Detected	Not Detected
	2,895,997 - 2,896,192	MI0001267	gga-mir-205a	Detected	N/A	Detected
27	3,553,047 - 3,553,241	MI0001268	gga-mir-196-1	Detected	N/A	Detected
28	4,435,975 - 4,436,169	MI0001269	gga-mir-7-3	Detected	N/A	Detected
	2,709,312 - 2,709,499	MI0003694	gga-mir-9-1	Detected	Detected	Detected
	1,784,344 - 1,784,517	MI0003711	gga-mir-211	Not Detected	N/A	Not Detected
Z	28,037,824 - 28,038,002	MI0001270	gga-mir-101	Detected	N/A	Detected
	39,554,716 - 39,554,924	MI0001272	gga-mir-7-1	Detected	N/A	Detected

Appendix 1 continued

	41,157,356 - 41,157,541	MI0001273	gga-mir-23b	Detected	N/A	Detected
	41,157,592 - 41,157,788	MI0001274	gga-mir-27b	Detected	N/A	Not Detected
	41,158,125 - 41,158,292	MI0001275	gga-mir-24	Detected	N/A	Not Detected
	71,882,121 - 71,882,314	MI0001276	gga-mir-31	Detected	N/A	Not Detected
	649,287 - 649,463	MI0001277	gga-mir-122a-1	Detected	N/A	Detected
	59,286,265 - 59,286,451	MI0001283	gga-mir-9-2	Detected	Detected	Detected
	16,040,563 - 16,040,748	MI0003715	gga-mir-449	Not Detected	N/A	Not Detected
Un_Random	38,163,771 - 38,163,980	MI0001279	gga-mir-7b	Detected	N/A	Detected
	12,066,746 - 12,066,922	MI0001280	gga-mir-122a-2	Detected	N/A	Detected
	27,776,406 - 27,776,613	MI0001282	gga-mir-196-3	Not Detected	N/A	Not Detected
	24,327,374 - 24,327,555	MI0005000	gga-mir-757-2	Not Detected	N/A	Not Detected
				Special Case N		
Not Validated		MI0001253	gga-mir-124b-2			
		MI0001271	gga-mir-204-1			
		MI0001278	gga-mir-183			
No Info		MI0001284	gga-mir-218-3			
10_random	11,844 - 12,013	MI0003696	gga-mir-147-1	Not Detected	N/A	Detected

ProMir-g can detect 78 pre-miRNAs out of 146 tested pre-miRNAs.

The 146 tested pre-miRNAs are divided into 88 miRNA groups. We can only use ERPIN to predict 22 groups out of those 88 groups. Because of this limitation, we can only test 58 pre-miRNA using ERPIN. The test resulted in 54 pre-miRNAs detection by ERPIN

Mir-abela detect 84 pre-miRNAs out of 146 tested pre-miRNAs.

Appendix 2 Negative Data Results using ProMir, ERPIN, and miR-abela

Gene Name	Position	ProMir II	ERPIN	miR-abela
DRG1 (Developmentally Regulated GTP binding protein)	241-480	None	None	None
ZP3 (Zona Pellucida glycoprotein 3) Sperm Receptor	421-660	None	None	None
RGMA RGM domain family, member A	361-600	None	None	None
ND2 NADH dehydrogenase subunit 2	121-360	None	None	None
ND4L NADH dehydrogenase subunit 4L	1-240	None	None	None
ND4 NADH dehydrogenase subunit 4	181-420	None	None	None
ND5 NADH dehydrogenase subunit 5	301-540	None	None	None
CYTB cytochrome b	1-240	None	None	None
ND6 NADH dehydrogenase subunit 6	61-300	None	None	None
COX1 cytochrome c oxidase subunit I	241-480	None	None	None
ND3 NADH dehydrogenase subunit 3	61-300	None	None	None
COX3 cytochrome c oxidase subunit III	1-240	None	None	None
ND1 NADH dehydrogenase subunit 1	361-600	None	None	None
COX2 cytochrome c oxidase subunit II	61-300	None	None	None
CPT1 carnitine palmitoyltransferase I	121-360	None	None	None
DAG1 dystroglycan	1-240	None	None	None
RCJMB04_12p16 ecotropic viral integration site 2A	121-360	None	None	None
RCJMB04_35d23 chromatin modifying protein 2A	1-240	None	None	None
RCJMB04_16n19 hypothetical protein LOC769534	181-420	None	None	None
ROR2 receptor tyrosine kinase-like orphan receptor 2	61-300	None	None	None
SIX2 sine oculis homeobox homolog 2	481-720	None	None	None
BLB1 MHC class II antigen B-F minor heavy chain	541-780	None	None	None
BF1 MHC class I antigen B-F minor heavy chain	961-1200	None	None	None
XKR8 XK, Kell blood group complex subunit-related family, member 8	661-900	None	None	None
RCJMB04_2f9 selenoprotein P, plasma, 1	301-540	None	None	None
ERBB2IP erbb2 interacting protein	1-240	None	None	None
GPR98 G protein-coupled receptor 98	1501-1740	None	None	None
RCJMB04_38d18 similar to topoisomerase 1-binding RING finger	721-960	None	None	None
ATP5A1W ATP synthase subunit alpha	421-660	None	None	None
RCJMB04_4o21 Nedd4 binding protein 3	781-1020	None	None	None
RFXDC2 regulatory factor X domain containing 2	301-540	None	None	None
POU3F1 POU domain, class 3, transcription factor 1	241-480	None	None	None

Appendix 2 continued

RCJMB04_33n14 NADH dehydrogenase (ubiquinone) flavoprotein 1, 51kDa	601-840	None	None	None
CXCR7 chemokine (C-X-C motif) receptor 7	61-300	None	None	None
CPS1 carbamoyl-phosphate synthetase 1, mitochondrial	421-660	None	None	None
PRLHR prolactin releasing hormone receptor	1-240	None	None	None
ADRA2A adrenergic, alpha-2A-, receptor	361-600	None	None	None
RCJMB04_5g20 phosphoglycerate mutase 1 (brain)	661-900	None	None	None
PPYR1 pancreatic polypeptide receptor 1	1-240	None	None	None
WNT8B wingless-type MMTV integration site family, member 8B	181-420	None	None	None
ARF6 ADP-ribosylation factor 6	1-240	None	None	None
RCJMB04_33f3 serologically defined colon cancer antigen 1	241-480	None	None	None
RCJMB04_8o12 zinc finger and BTB domain containing 1	841-1080	None	None	None
JAG2 jagged 2	301-540	None	None	None
BDKRB2 bradykinin receptor B2	1-240	None	None	None
TSHR thyroid stimulating hormone receptor	241-480	None	None	None
RCJMB04_1f2 ADP-ribosylation-like factor 6 interacting protein 4	301-540	None	None	None
RCJMB04_30k7 strawberry notch homolog 1	601-840	None	None	None
BTBD3 BTB (POZ) domain containing 3	421-660	None	None	None
DNASE1L2 deoxyribonuclease I-like 2	1-240	None	None	None
TMEM8 transmembrane protein 8 (five membrane-spanning domains)	601-840	None	None	None
RCJMB04_20k18 phosphomannomutase 2	481-720	None	None	None
RCJMB04_20f12 glutamyl-tRNA synthetase 2, mitochondrial (putative)	601-840	None	None	None
SSTR5 somatostatin receptor 5	1-240	None	None	None
RCJMB04_6k9 transmembrane protein 11	541-780	None	None	None
RCJMB04_2o8 ras homolog gene family, member T2	361-600	None	None	None
SPRY4 sprouty homolog 4	121-360	None	None	None
FNIP1 folliculin interacting protein 1	1021-1260	None	None	None
DRD1 dopamine receptor D1	1201-1440	None	None	None
SH3PXD2B SH3 and PX domains 2B	1-240	None	None	None
PCDHGC5 protocadherin gamma subfamily C, 5	721-960	None	None	None
DOCK2 dedicator of cytokinesis 2	121-360	None	None	None
RCJMB04_27b6 zinc finger, matrin type 2	541-780	None	None	None
RCJMB04_32j5 small glutamine-rich tetratricopeptide repeat (TPR)-containing, alpha	1801-2040	None	None	None
RCJMB04_18m15 similar to BC027088 protein	301-540	None	None	None
RCJMB04_20f16 family with sequence similarity 76, member A	181-420	None	None	None

Appendix 2 continued

RCJMB04_10n8 embryonic ectoderm development	601-840	None	None	None
SEC61A2 Sec61 alpha 2 subunit (S. cerevisiae)	1381-1620	None	None	None
NUDT5 nudix (nucleoside diphosphate linked moiety X)-type motif 5	1-240	None	None	None
CDC123 cell division cycle 123 homolog (S. cerevisiae)	901-1140	None	None	None
RCJMB04_9e6 DnaJ (Hsp40) homolog, subfamily B, member 14	1021-1260	None	None	None
RCJMB04_2h11 H2A histone family, member V	601-840	None	None	None
S100A9 S100 calcium binding protein A9	1-240	None	None	None
RCJMB04_5b12 ubiquilin 4	1501-1740	None	None	None
RCJMB04_28k7 TAP binding protein-like	301-540	None	None	None
QDPR quinoid dihydropteridine reductase	601-840	None	None	None
RCJMB04_3b8 protein phosphatase 2A activator, regulatory subunit 4	1021-1260	None	None	None
RCJMB04_23i8 zinc finger CCCH-type, antiviral 1	661-900	None	None	None
TOMM20 translocase of outer mitochondrial membrane 20 homolog (yeast)	301-540	None	None	None
HOXB5 homeobox B5	1201-1440	None	None	None
CDCA8 cell division cycle associated 8	1-240	None	None	None
ENSA endosulfine alpha	721-960	None	None	None
GOLPH3L golgi phosphoprotein 3-like	601-840	None	None	None
MSTO1 misato homolog 1 (Drosophila)	661-900	None	None	None
DAP3 death associated protein 3	421-660	None	None	None
ASH1L ash1 (absent, small, or homeotic)-like (Drosophila)	721-960	None	None	None
CYP17A1 cytochrome P450, family 17, subfamily A, polypeptide 1	661:900	None	None	None
CYP8B cytochrome P450, family 8, subfamily B	1021-1260	None	None	None
ZNF403 zinc finger protein 403	241-480	None	None	None
TCP11L2 t-complex 11 (mouse)-like 2	601-840	None	None	None
GYG1 glycogenin 1	901-1140	None	None	None
HPS3 Hermansky-Pudlak syndrome 3	661-900	None	None	None
SELT selenoprotein T	601-840	None	None	None
GPR87 G protein-coupled receptor 87	181-420	None	None	None
AADAC arylacetamide deacetylase (esterase)	301-540	None	None	None
MME membrane metallo-endopeptidase	601-840	None	None	None
PLCH1 phospholipase C, eta 1	1321-1560	None	None	None
SLC33A1 solute carrier family 33 (acetyl-CoA transporter), member 1	901-1140	None	None	None
GMPS guanine monphosphate synthetase	601-840	None	None	None
SSR3 signal sequence receptor, gamma (translocon-associated protein gamma)	1-240	None	None	None

Appendix 2 continued

>chromosome:WASHUC2:1:170425725:170425951:-1 (rRNA)		None	None	None
>chromosome:WASHUC2:2:44498935:44499153:-1		None	None	None
>chromosome:WASHUC2:6:9697078:9697296:1		None	None	None
>chromosome:WASHUC2:9:1895011:1895229:-1		None	None	None
>chromosome:WASHUC2:9:1892308:1892526:-1		None	None	None
>chromosome:WASHUC2:Un_random:30104724:30104944:1		None	None	None
>chromosome:WASHUC2:1:136134114:136134332:1		None	None	None
>chromosome:WASHUC2:5:41411741:41411958:1		None	None	None
>chromosome:WASHUC2:23:1575045:1575275:-1		None	None	None
>chromosome:WASHUC2:9:4405951:4406169:1		None	None	None
>chromosome:WASHUC2:1:104486616:104486868:-1		None	None	None
>chromosome:WASHUC2:Un_random:8223876:8224128:1		None	None	None
>chromosome:WASHUC2:9:1897191:1897409:-1		None	None	None
>chromosome:WASHUC2:23:962429:962645:1		None	None	None
>chromosome:WASHUC2:1:46950816:46950987:1 (tRNA)		None	None	None
>chromosome:WASHUC2:1:48519106:48519277:1		None	None	None
>chromosome:WASHUC2:1:87492901:87493074:-1		None	None	None
>chromosome:WASHUC2:1:87494218:87494391:1		None	None	None
>chromosome:WASHUC2:1:87494699:87494870:-1		None	None	None
>chromosome:WASHUC2:1:87496185:87496358:1		None	None	None
>chromosome:WASHUC2:1:89105779:89105952:1		None	None	None
>chromosome:WASHUC2:1:123939276:123939448:1		None	None	None
>chromosome:WASHUC2:1:142239652:142239824:-1		None	None	None
>chromosome:WASHUC2:1:150119479:150119651:-1		None	None	None
>chromosome:WASHUC2:1:152583919:152584090:1		None	None	None
>chromosome:WASHUC2:1:153533676:153533857:-1		None	None	None
>chromosome:WASHUC2:1:169455877:169456052:1		None	None	None
>chromosome:WASHUC2:1:179369242:179369414:1		None	None	None
>chromosome:WASHUC2:10:20307018:20307189:-1		None	None	None
>chromosome:WASHUC2:11:1637786:1637956:1		None	None	None
>chromosome:WASHUC2:11:10726389:10726562:1		None	None	None
>chromosome:WASHUC2:12:10326779:10326961:1		None	None	None

Appendix 2 continued

>chromosome:WASHUC2:14:8887673:8887854:1		None	None	None
>chromosome:WASHUC2:15:4537870:4538041:1		None	None	None
>chromosome:WASHUC2:16:162661:162834:1		None	None	None
>chromosome:WASHUC2:16:161914:162095:-1		None	None	None
>chromosome:WASHUC2:16:147237:147409:1		None	None	None
>chromosome:WASHUC2:16:145107:145278:1		None	None	None
>chromosome:WASHUC2:16:141682:141854:-1		None	None	None
>chromosome:WASHUC2:16:121783:121955:1		None	None	None
>chromosome:WASHUC2:17:1694385:1694557:1		None	None	None
>chromosome:WASHUC2:18:3354874:3355045:-1		None	None	None
>chromosome:WASHUC2:18:6413125:6413296:1		None	None	None
>chromosome:WASHUC2:2:17682766:17682939:1		None	None	None
>chromosome:WASHUC2:2:21440395:21440568:-1		None	None	None
>chromosome:WASHUC2:2:62134854:62135043:-1		None	None	None
>chromosome:WASHUC2:2:92062952:92063124:-1		None	None	None
>chromosome:WASHUC2:2:96038061:96038233:-1		None	None	None
>chromosome:WASHUC2:2:119281892:119282081:-1		None	None	None
>chromosome:WASHUC2:2:119282658:119282830:-1		None	None	None
>chromosome:WASHUC2:2:119283008:119283197:1		None	None	None
>chromosome:WASHUC2:2:131256074:131256255:-1		None	None	None
>chromosome:WASHUC2:2:153829982:153830154:1		None	None	None
>chromosome:WASHUC2:2:153683664:153683836:1		None	None	None
>chromosome:WASHUC2:2:149662183:149662355:-1		None	None	None
>chromosome:WASHUC2:2:148210803:148210981:1		None	None	None
>chromosome:WASHUC2:2:145457517:145457698:-1		None	None	None
>chromosome:WASHUC2:2:143572890:143573062:-1		None	None	None
>chromosome:WASHUC2:20:2511099:2511270:-1		None	None	None
>chromosome:WASHUC2:27:2341498:2341679:-1		None	None	None
>chromosome:WASHUC2:28:4466049:4466221:1		None	None	None
>chromosome:WASHUC2:3:19206925:19207097:-1		None	None	None
>chromosome:WASHUC2:3:47787339:47787521:1		None	None	None
>chromosome:WASHUC2:3:50069956:50070126:-1		None	None	None

Appendix 2 continued

>chromosome:WASHUC2:3:74405789:74405961:-1		None	None	None
>chromosome:WASHUC2:3:79459192:79459363:1		None	None	None
>chromosome:WASHUC2:3:83512599:83512770:-1		None	None	None
>chromosome:WASHUC2:3:95102927:95103108:-1		None	None	None
>chromosome:WASHUC2:3:99438779:99438950:1		None	None	None
>chromosome:WASHUC2:3:110215916:110216088:-1		None	None	None
>chromosome:WASHUC2:3:110217278:110217449:1		None	None	None
>chromosome:WASHUC2:4:45355672:45355843:1		None	None	None
>chromosome:WASHUC2:5:17258216:17258388:1		None	None	None
>chromosome:WASHUC2:5:52074722:52074895:1		None	None	None
>chromosome:WASHUC2:5:57853238:57853410:1		None	None	None
>chromosome:WASHUC2:5:58901909:58902080:-1		None	None	None
>chromosome:WASHUC2:6:18106467:18106638:1		None	None	None
>chromosome:WASHUC2:6:33388741:33388912:1		None	None	None
>chromosome:WASHUC2:7:4192971:4193153:1		None	None	None
>chromosome:WASHUC2:7:6794615:6794786:1		None	None	None
>chromosome:WASHUC2:7:6798357:6798538:-1		None	None	None
>chromosome:WASHUC2:7:6800690:6800871:1		None	None	None
>chromosome:WASHUC2:7:6813478:6813659:-1		None	None	None
>chromosome:WASHUC2:7:6815754:6815935:-1		None	None	None
>chromosome:WASHUC2:7:6816497:6816678:1		None	None	None
>chromosome:WASHUC2:7:37679897:37680067:-1		None	None	None
>chromosome:WASHUC2:8:26034025:26034195:1		None	None	None
>chromosome:WASHUC2:9:23003789:23003962:-1		None	None	None
>chromosome:WASHUC2:MT:15989:16157:1		None	None	None
>chromosome:WASHUC2:MT:12950:13120:1		None	None	None
>chromosome:WASHUC2:1:33622572:33622796:-1 (snRNA)		None	None	None
>chromosome:WASHUC2:1:114732509:114732708:1		None	None	None
>chromosome:WASHUC2:1:192763619:192763864:1		None	None	None
>chromosome:WASHUC2:11:5326386:5326592:1		None	None	None
>chromosome:WASHUC2:15:9656919:9657159:1		None	None	None
>chromosome:WASHUC2:17:7891746:7891971:-1		None	None	None

Appendix 2 continued

>chromosome:WASHUC2:18:3449237:3449443:-1		None	None	None
>chromosome:WASHUC2:2:37104950:37105201:-1		None	None	None
>chromosome:WASHUC2:20:1534620:1534883:-1		None	None	None
>chromosome:WASHUC2:28:2332386:2332592:1		None	None	None

ERPIN tested on all 21 available miRNA groups

Appendix 3 Detail Results of ProMir, ERPIN, miR-abela, and miR-Explore.

Chromosome #	Position	Accession	ID	ProMir II	ERPIN	miR-abela	miR-Explore
1	3,236,279 - 3,236,467	MI0001166	gga-mir-29a	Detected	Detected	Detected	Detected
	3,235,262 - 3,235,442	MI0001167	gga-mir-29b-1	Detected	Detected	Detected	Detected
	34,895,637 - 34,895,820	MI0001168	gga-let-7i	Detected	Detected	Detected	Detected
	48,192,609 - 48,192,808	MI0001169	gga-mir-135a-2	Detected	N/A	Detected	Detected
	51,372,232 - 51,372,400	MI0001170	gga-mir-33	Not Detected	N/A	Not Detected	Detected
	73,421,222 - 73,421,397	MI0001171	gga-let-7a-3	Detected	Detected	Detected	Detected
	73,422,051 - 73,422,235	MI0001172	gga-let-7b	Not Detected	Detected	Not Detected	Detected
	102,424,283 - 102,424,463	MI0001173	gga-mir-99a	Detected	Detected	Detected	Detected
	102,425,036 - 102,425,219	MI0001174	gga-let-7c	Detected	Detected	Detected	Detected
	102,457,597 - 102,457,786	MI0001175	gga-mir-125b	Detected	Detected	Detected	Detected
	105,930,163 - 105,930,325	MI0001176	gga-mir-155	Detected	N/A	Detected	Detected
	114,215,977 - 114,216,174	MI0001177	gga-mir-222	Not Detected	N/A	Not Detected	Detected
	114,218,372 - 114,218,569	MI0001177	gga-mir-222	Detected	N/A	Not Detected	Detected
	114,218,876 - 114,219,074	MI0001178	gga-mir-221	Detected	N/A	Detected	Detected
	152,248,020 - 152,248,197	MI0001179	gga-mir-92	Not Detected	Detected	Not Detected	Not Detected
	152,248,133 - 152,248,319	MI0001180	gga-mir-19b	Detected	N/A	Not Detected	Not Detected
	152,248,256 - 152,248,453	MI0001181	gga-mir-20a	Not Detected	Detected	Detected	Detected
	152,248,442 - 152,248,572	MI0001182	gga-mir-19a	Not Detected	N/A	Not Detected	Not Detected
	152,248,576 - 152,248,768	MI0001183	gga-mir-18a	Not Detected	Detected	Not Detected	Detected
	152,248,731 - 152,248,915	MI0001184	gga-mir-17	Not Detected	Detected	Detected	Detected
	173,700,301 - 173,700,484	MI0001185	gga-mir-16-1	Not Detected	Detected	Not Detected	Not Detected
	173,700,443 - 173,700,625	MI0001186	gga-mir-15a	Not Detected	Detected	Not Detected	Detected
	1,147,470 - 1,147,667	MI0001238	gga-mir-205b	Detected	N/A	Detected	Detected
	59,948,651 - 59,948,843	MI0003709	gga-mir-490	Not Detected	N/A	Not Detected	Detected
2	4,467,466 - 4,467,642	MI0001187	gga-mir-26a	Not Detected	N/A	Detected	Detected
	8,765,637 - 8,765,823	MI0001188	gga-mir-153	Not Detected	N/A	Detected	Detected
	32,053,493 - 32,053,660	MI0001189	gga-mir-148a	Not Detected	N/A	Not Detected	Detected
	32,586,099 - 32,586,292	MI0001190	gga-mir-196-2	Not Detected	N/A	Not Detected	Detected
	40,745,098 - 40,745,293	MI0001191	gga-mir-138-1	Not Detected	N/A	Not Detected	Detected
	45,549,126 - 45,549,309	MI0001192	gga-mir-128-2	Not Detected	N/A	Detected	Detected

Appendix 3 continued

	85,892,420 - 85,892,605	MI0001193	gga-mir-187	Detected	N/A	Detected	Detected
	86,506,401 - 86,506,570	MI0001194	gga-mir-32	Not Detected	N/A	Not Detected	Detected
	105,670,307 - 105,670,493	MI0001195	gga-mir-133a-1	Not Detected	Detected	Detected	Detected
	105,673,433 - 105,673,617	MI0001196	gga-mir-1a-2	Not Detected	Detected	Not Detected	Detected
	118,524,107 - 118,524,302	MI0001197	gga-mir-124a	Not Detected	Detected	Detected	Detected
	148,337,213 - 148,337,376	MI0001198	gga-mir-30d	Not Detected	Detected	Not Detected	Not Detected
	148,331,548 - 148,331,734	MI0001199	gga-mir-30b	Not Detected	Detected	Not Detected	Detected
	23,068,827 - 23,069,010	MI0003708	gga-mir-489	Not Detected	N/A	Not Detected	Detected
	21,671,911 - 21,672,079	MI0003713	gga-mir-466	Not Detected	N/A	Not Detected	Detected
	3,583,640 - 3,583,829	MI0004998	gga-mir-460	Not Detected	N/A	Not Detected	Detected
	78,310,794 - 78,310,975	MI0004999	gga-mir-757-1	Not Detected	N/A	Not Detected	Detected
3	282,757 - 282,962	MI0001200	gga-mir-216	Not Detected	N/A	Not Detected	Not Detected
	280,089 - 280,295	MI0001201	gga-mir-217	Detected	N/A	Not Detected	Detected
	19,924,437 - 19,924,611	MI0001202	gga-mir-194	Not Detected	N/A	Detected	Detected
	19,924,743 - 19,924,947	MI0001203	gga-mir-215	Not Detected	N/A	Detected	Detected
	85,102,189 - 85,102,360	MI0001204	gga-mir-30a	Detected	Detected	Detected	Detected
	85,126,803 - 85,126,974	MI0001205	gga-mir-30c-2	Not Detected	Detected	Not Detected	Detected
	110,384,883 - 110,385,066	MI0001206	gga-mir-133b	Not Detected	Detected	Detected	Detected
	110,390,389 - 110,390,564	MI0001207	gga-mir-206	Not Detected	Detected	Not Detected	Detected
	32,679,660 - 32,679,871	MI0004997	gga-mir-456	Not Detected	N/A	Not Detected	Detected
4	232,899 - 233,098	MI0001208	gga-mir-223	Detected	Detected	Detected	Detected
	3,970,178 - 3,970,361	MI0001209	gga-mir-18b	Not Detected	Detected	Detected	Detected
	3,970,309 - 3,970,489	MI0001210	gga-mir-106	Detected	Not Detected	Detected	Detected
	58,651,829 - 58,651,995	MI0001211	gga-mir-302a	Detected	Detected	Detected	Detected
	77,774,648 - 77,774,856	MI0001212	gga-mir-218-1	Detected	N/A	Detected	Detected
	91,906,839 - 91,907,021	MI0001213	gga-mir-103-2	Not Detected	N/A	Not Detected	Detected
	3,969,997 - 3,970,181	MI0001517	gga-mir-20b	Not Detected	Detected	Not Detected	Detected
	58,651,264 - 58,651,435	MI0003700	gga-mir-302b	Detected	Not Detected	Detected	Detected
	58,651,526 - 58,651,690	MI0003701	gga-mir-302c	Detected	Not Detected	Detected	Detected
	58,652,164 - 58,652,332	MI0003702	gga-mir-302d	Detected	Not Detected	Detected	Detected
	65,844,645 - 65,844,817	MI0003706	gga-mir-383	Not Detected	N/A	Not Detected	Detected
	58,652,310 - 58,652,472	MI0003712	gga-mir-367	Detected	N/A	Detected	Detected

Appendix 3 continued

5	53,206,764 - 53,206,961	MI0001214	gga-mir-203	Not Detected	Detected	Detected	Detected
6	20,487,914 - 20,488,094	MI0001215	gga-mir-107	Not Detected	N/A	Not Detected	Detected
	24,570,010 - 24,570,214	MI0003695	gga-mir-146b	Detected	N/A	Detected	Not Detected
	22,813,018 - 22,813,206	MI0003699	gga-mir-202	Detected	N/A	Detected	Detected
7	17,388,998 - 17,389,207	MI0001216	gga-mir-10b	Not Detected	Detected	Not Detected	Detected
	32,228,100 - 32,228,281	MI0001217	gga-mir-128-1	Not Detected	N/A	Detected	Detected
	23,901,074 - 23,901,238	MI0003705	gga-mir-375	Not Detected	N/A	Detected	Detected
8	2,001,511 - 2,001,714	MI0001218	gga-mir-181a-1	Detected	N/A	Detected	Detected
	2,001,700 - 2,001,888	MI0001219	gga-mir-181b-1	Detected	N/A	Not Detected	Detected
	4,732,723 - 4,732,930	MI0001220	gga-mir-199-2	Not Detected	N/A	Detected	Detected
	13,210,143 - 13,210,338	MI0001221	gga-mir-137	Not Detected	N/A	Detected	Detected
9	23,742,741 - 23,742,934	MI0001222	gga-mir-16-2	Detected	Detected	Detected	Detected
	23,742,916 - 23,743,106	MI0001223	gga-mir-15b	Detected	Detected	Not Detected	Detected
10	5,209,674 - 5,209,858	MI0001224	gga-mir-190	Not Detected	N/A	Detected	Detected
	6,651,224 - 6,651,424	MI0001225	gga-mir-204-2	Detected	N/A	Detected	
	14,823,475 - 14,823,673	MI0001226	gga-mir-7-2	Not Detected	N/A	Not Detected	Detected
	22,146,195 - 22,146,368	MI0001227	gga-mir-184	Detected	N/A	Detected	Detected
	12,334,872 - 12,335,041	MI0003697	gga-mir-147-2	Not Detected	N/A	Detected	Detected
11	2,023,904 - 2,024,086	MI0001228	gga-mir-138-2	Not Detected	N/A	Not Detected	Detected
	21,030,591 - 21,030,785	MI0001229	gga-mir-140	Detected	Detected	Detected	Detected
12	2,809,028 - 2,809,210	MI0001230	gga-let-7g	Detected	Detected	Detected	Detected
	2,830,692 - 2,830,879	MI0001231	gga-mir-135a-1	Not Detected	N/A	Not Detected	Not Detected
	6,301,402 - 6,301,604	MI0001232	gga-let-7d	Detected	Detected	Detected	Detected
	6,302,447 - 6,302,633	MI0001233	gga-let-7f	Not Detected	Detected	Not Detected	Detected
	6,302,861 - 6,303,050	MI0001234	gga-let-7a-1	Detected	Detected	Not Detected	Detected

Appendix 3 continued

13	7,555,543 - 7,555,741	MI0001235	gga-mir-146a	Detected	N/A	Detected	Not Detected
	4,449,192 - 4,449,369	MI0001236	gga-mir-103-1	Detected	N/A	Detected	Detected
	4,322,810 - 4,323,004	MI0001237	gga-mir-218-2	Not Detected	N/A	Detected	Detected
14	759,403 - 759,585	MI0003698	gga-mir-193	Not Detected	N/A	Not Detected	Detected
	764,221 - 764,405	MI0003703	gga-mir-365-1	Detected	N/A	Not Detected	Detected
15	398,670 - 398,846	MI0001239	gga-mir-130b	Not Detected	N/A	Detected	Detected
	406,263 - 406,455	MI0001240	gga-mir-301	Detected	N/A	Not Detected	Not Detected
	408,349 - 408,531	MI0001241	gga-mir-130a	Detected	N/A	Not Detected	Detected
17	10,220,087 - 10,220,271	MI0001242	gga-mir-181b-2	Detected	N/A	Detected	Detected
	10,218,447 - 10,218,637	MI0001243	gga-mir-181a-2	Not Detected	N/A	Detected	Detected
	8,431,692 - 8,431,875	MI0001244	gga-mir-126	Detected	N/A	Not Detected	Detected
	5,667,100 - 5,667,293	MI0001245	gga-mir-199-1	Detected	N/A	Detected	Detected
	5,577,767 - 5,577,951	MI0001246	gga-mir-219	Detected	N/A	Detected	Detected
	5,339,651 - 5,339,836	MI0003707	gga-mir-455	Detected	N/A	Detected	Detected
18	6,437,246 - 6,437,441	MI0003704	gga-mir-365-2	Detected	N/A	Detected	Detected
19	496,933 - 497,120	MI0001281	gga-mir-142	Detected	N/A	Detected	Detected
	7,322,022 - 7,322,218	MI0004994	gga-mir-21	Detected	N/A	Detected	Detected
	5,823,918 - 5,824,086	MI0004995	gga-mir-451	Detected	N/A	Detected	Detected
	5,824,073 - 5,824,257	MI0004996	gga-mir-144	Detected	N/A	Not Detected	Detected
20	8,107,781 - 8,107,951	MI0001247	gga-mir-1a-1	Detected	Detected	Detected	Detected
	8,119,004 - 8,119,199	MI0001248	gga-mir-133a-2	Detected	Detected	Detected	Detected
	2,599,284 - 2,599,474	MI0003710	gga-mir-499	Detected	N/A	Detected	Detected
21	2,583,267 - 2,583,453	MI0001249	gga-mir-200a	Detected	N/A	Detected	Detected
	2,585,592 - 2,585,776	MI0001250	gga-mir-200b	Detected	N/A	Not Detected	Detected
	3,251,464 - 3,251,672	MI0001251	gga-mir-34a	Not Detected	Detected	Not Detected	Not Detected
	2,580,762 - 2,580,945	MI0003714	gga-mir-429	Not Detected	N/A	Not Detected	Detected

Appendix 3 continued

23	2,510,281 - 2,510,473	MI0001252	gga-mir-124b-1	Detected	Detected	Detected	Detected
	4,663,862 - 4,664,025	MI0001254	gga-mir-1b	Detected	Detected	Not Detected	Detected
	4,664,001 - 4,664,179	MI0001255	gga-mir-133c	Not Detected	Detected	Not Detected	Detected
	5,248,364 - 5,248,559	MI0001256	gga-mir-30e	Detected	Detected	Detected	Detected
	5,249,587 - 5,249,775	MI0001257	gga-mir-30c-1	Not Detected	Detected	Detected	Detected
24	3,372,844 - 3,373,023	MI0001258	gga-mir-100	Detected	Detected	Detected	Detected
	3,380,943 - 3,381,114	MI0001259	gga-let-7a-2	Detected	Detected	Detected	Detected
	5,684,850 - 5,685,033	MI0001260	gga-mir-34b	Detected	Detected	Detected	Detected
	5,685,587 - 5,685,760	MI0001261	gga-mir-34c	Detected	Detected	Detected	Detected
26	1,442,647 - 1,442,829	MI0001262	gga-let-7j	Not Detected	Detected	Not Detected	Detected
	1,442,847 - 1,443,029	MI0001263	gga-let-7k	Detected	Detected	Not Detected	Detected
	1,925,892 - 1,926,087	MI0001264	gga-mir-135a-3	Not Detected	N/A	Not Detected	Not Detected
	2,511,608 - 2,511,796	MI0001265	gga-mir-29c	Not Detected	Detected	Not Detected	Detected
	2,512,519 - 2,512,698	MI0001266	gga-mir-29b-2	Not Detected	Detected	Not Detected	Detected
	2,895,997 - 2,896,192	MI0001267	gga-mir-205a	Detected	N/A	Detected	Detected
27	3,553,047 - 3,553,241	MI0001268	gga-mir-196-1	Detected	N/A	Detected	Detected
28	4,435,975 - 4,436,169	MI0001269	gga-mir-7-3	Detected	N/A	Detected	Not Detected
	2,709,312 - 2,709,499	MI0003694	gga-mir-9-1	Detected	Detected	Detected	Not Detected
	1,784,344 - 1,784,517	MI0003711	gga-mir-211	Not Detected	N/A	Not Detected	
Z	28,037,824 - 28,038,002	MI0001270	gga-mir-101	Detected	N/A	Detected	Detected
	39,554,716 - 39,554,924	MI0001272	gga-mir-7-1	Detected	N/A	Detected	Detected
	41,157,356 - 41,157,541	MI0001273	gga-mir-23b	Detected	N/A	Detected	Detected
	41,157,592 - 41,157,788	MI0001274	gga-mir-27b	Detected	N/A	Not Detected	Detected
	41,158,125 - 41,158,292	MI0001275	gga-mir-24	Detected	N/A	Not Detected	Detected
	71,882,121 - 71,882,314	MI0001276	gga-mir-31	Detected	N/A	Not Detected	Detected
	649,287 - 649,463	MI0001277	gga-mir-122a-1	Detected	N/A	Detected	Detected
	59,286,265 - 59,286,451	MI0001283	gga-mir-9-2	Detected	Detected	Detected	Not Detected
	16,040,563 - 16,040,748	MI0003715	gga-mir-449	Not Detected	N/A	Not Detected	Not Detected

Appendix 3 continued

Un_Random	38,163,771 - 38,163,980	MI0001279	gga-mir-7b	Detected	N/A	Detected	Detected
	12,066,746 - 12,066,922	MI0001280	gga-mir-122a-2	Detected	N/A	Detected	Detected
	27,776,406 - 27,776,613	MI0001282	gga-mir-196-3	Not Detected	N/A	Not Detected	Detected
	24,327,374 - 24,327,555	MI0005000	gga-mir-757-2	Not Detected	N/A	Not Detected	Detected
				Special Case N			
Not Validated		MI0001253	gga-mir-124b-2				
		MI0001271	gga-mir-204-1				
		MI0001278	gga-mir-183				
No Info		MI0001284	gga-mir-218-3				
10_random	11,844 - 12,013	MI0003696	gga-mir-147-1	Not Detected	N/A	Detected	Detected

ProMir II result
Not Detected = 0.465753425 Detected = 0.534246575

ERPIN result
Total of 88 groups where 22 groups availabel in ERPIN
Of all groups that is available in ERPIN, the program can detect 93% known
N/A : miRNA group is not available in ERPIN
Not Detected : miRNA group is available, but ERPIN can not detect the miRNA in the input

miR-abela
result
Not Detected : 62/146 = 42.47%
Detected : 84/146 = 57.53%

miR-Finder
Not Detected 16/146 = 10.96%
Detected 130/146 = 89.04%