國立交通大學

生物科技研究所

碩士論文

miRTar: 快速而準確預測 microRNA target 之網站工具

miRTar: a fast and efficient microRNA target prediction web server

研究生：許勝達

指導教授：黃憲達　博士

中華民國九十五年六月

miRTar: 快速而準確預測 microRNA target 之網站工具

miRTar: a fast and efficient microRNA target prediction web server

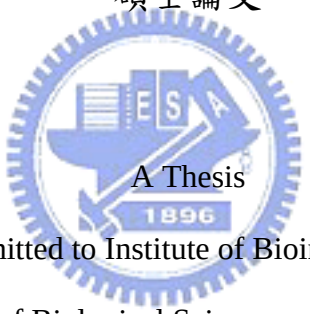
研 究 生：許勝達　　Student：Sheng-Da Hsu

指導教授：黃憲達 博士　　Advisor：Dr. Hsien-Da Huang

國立交通大學

生物科技研究所

碩士論文



A Thesis

Submitted to Institute of Bioinformatics

College of Biological Science and Technology

National Chiao Tung University

In partial Fulfillment of the Requirements

For the Degree of

Master

In

Biological Science and Technology

June 2006

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 五 年 六 月

miRTar: 快速而準確預測 microRNA target 之網站工具

學生：許勝達　　指導教授：黃憲達 博士

國立交通大學 生物科技研究所碩士班

中文摘要

目前已經有很多 microRNAs 被發現，而且也進一步透過實驗驗證。針對找尋 microRNA target，有許多的軟體被設計開發出來，例如 miRanda、RNAhybrid、和 TargetScan，但是，當使用這些軟體在大量的序列上，找尋 microRNA target 時是很耗時的，況且因為沒有很人性化的輸入查詢界面，對一個生物學家而言是非常不方便的。因此，為了幫助生物學家可以簡便的分析預測 microRNA target，一個網頁工具是必要的。本研究主要的貢獻為設計一網站工具-miRTar，提供更人性化的界面引導使用者輸入自訂的 microRNA 序列或 miRBase 的 microRNA ID，並在哺乳動物基因的保留序列上找尋出該 microRNA 的 target。另一方面，我們也會提供 mRNA 上面有 miRNA target site 的 RNA 二級結構相關分析。miRTar 設計的概念，主要是運用一個 dynamic programming 演算法的過濾步驟，再藉由 miRanda、RNAhybrid 和 TargetScan 來預測 miRNA targets。經過與 miRanda、RNAhybrid 和 TargetScan 的效能評比，miRTar 在效能方面有顯著的改進。miRTar 網站目前建置在 http://miRTar.mbc.nctu.edu.tw。

miRTar: a fast and efficient microRNA target prediction web server

Student：Sheng-Da Hsu　　Advisor：Dr . Hsien-Da Huang

Institute of Biological Science and Technology,
National Chiao Tung University

Abstract

Numerous microRNAs (miRNA) have been identified and experimentally validated. For identifying miRNA targets, a variety of programs such as miRanda, RNAhybrid and TargetScan have been developed. However these tools are time-consuming and inconvenient for biologists when predicting miRNA targets in a large sequence database. In order to perform the analysis in a more convenient manner, a web-based tool for identifying miRNA targets is crucial. This work presents an efficient web server, namely miRTar, which utilizes an intuitive interface that allows users to input a user-defined miRNA sequence or accession numbers of the miRBase for identifying miRNA targets against the conserved sequences of mRNAs of mammalian genes. Furthermore, this work also provides additional information about the RNA secondary structures of the mRNA containing the miRNA target site, which can be targeted by miRNAs. The miRTar utilizes a filtering strategy, which was implemented based on dynamic programming, just before applying miRanda, RNAhybrid and TargetScan for miRNA target prediction. By comparing the proposed web server to miRanda, RNAhybrid and TargetScan, miRTar performs remarkably more efficiently than those software. This prediction web server is now available at http://miRTar.mbc.nctu.edu.tw/.

致　謝

首先，我要感謝指導教授黃憲達博士在這兩年的日子當中對於我的細心指導，使得我可以在生物資訊這個領域內從無到有的學習到許多知識，也在學術研究上有顯著的進步及成長。

實驗室的學長們，謝謝你們對我的細心指導，實驗室的同學們，謝謝大家在這兩年內的互相幫忙及鼓勵，和大家一起討論的日子，是我成長的動力，實驗室內的點點滴滴更是美好的回憶。

最後，我要特別感謝我的家人給予我的支持，謝謝你們給予我的支持與鼓勵，才能讓我無後顧之憂的求學。

能夠順利完成碩士論文並取得碩士學位，是大家的指導、支持、與鼓勵，誠心的謝謝大家，將這份喜悅及成果與關心我的所有人一同分享。

國立交通大學 生物科技研究所

發現生物資訊實驗室 研究生 許勝達

謹誌於交通大學　2006 年六月

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

1.1.1 Non-coding RNA

The central dogma of molecular biology is based on the principle that the flow of genetic information travels from DNA to RNA and finally to the translation of proteins (Fig. 1.1). But in the recent years, a lot of non-protein coding RNAs (or noncoding RNAs, ncRNAs) have been discovered. What is known as noncoding RNA? Noncoding RNA refers to mRNA that is transcribed from DNA but not translated into protein. Rather than being junk DNA, noncoding RNA in fact play a critical role in regulating gene expression and so maintenance of more complex organism.

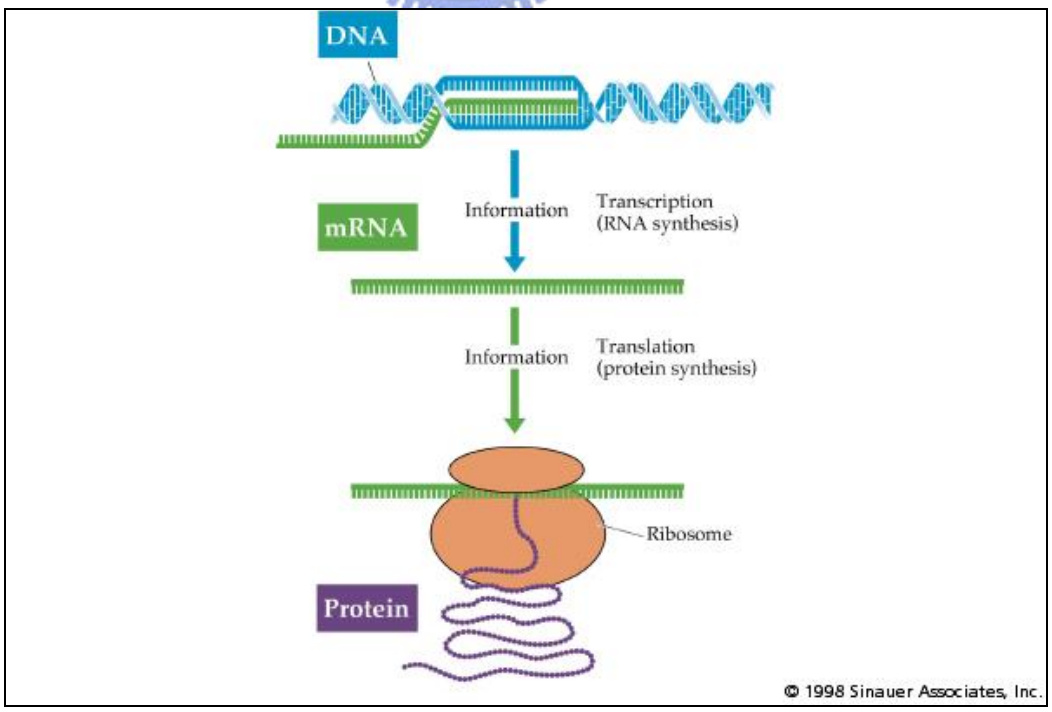


Figure 1.1 Central dogma of molecular biology.

Small non-coding RNAs can be divided into four subclass: microRNAs (miRNAs) [1-4], short interfering RNAs (siRNAs) [5, 6], tiny non-coding RNAs (tncRNAs) [7] and unique double-stranded RNAs called small modulatory RNAs (smRNAs) [8]. The characteristics of these classes have been reviewed in this report [9], as listed Table 1.1.

Subclass	Main characteristics
siRNAs	A class of double-stranded RNAs of 21–22 nucleotides in length, generated from dsRNAs. siRNAs silence genes by promoting the cleavage of mRNAs with exactly complementary sequences, or recruiting inhibitory proteins to, or directing the modification of, DNAs with exactly complementary sequences.
miRNAs	A class of 19–25-nucleotide, single-stranded RNAs that are encoded in the genomes of most multicellular organisms studied. Some are evolutionarily conserved and are developmentally regulated. They silence certain cellular genes at the stage of protein synthesis.
tncRNAs	A newly discovered class of short, 20–22-nucleotide RNAs that are encoded in the genome of C. elegans. They are not evolutionarily conserved, but some are developmentally regulated. Their function is still unknown.
smRNA	A short, dsRNA, identified earlier this year in mice, that allows the expression of neuron-specific genes only in adult neurons.

Table 1.1 The main characteristics of short RNAs [9].

MicroRNAs (miRNAs) constitute a large family of noncoding RNAs that function as guide molecules in diverse gene silencing pathways. This work focuses on the miRNAs and on what roles they play in the regulation of gene expression.

1.1.2 microRNAs

In 1993, two papers suggested that, in *C. elegans*, a small noncoding RNA, called *lin-4*, was responsible for regulating the expression of the *lin-14* gene through direct interaction with its mRNA [10, 11].

MicroRNAs (miRNAs) are small noncoding RNA molecules that are ~22 nts sequences capable of suppressing mRNA translation or mediate mRNA degradation by typically hybridizing to the 3'-untranslated regions (3'-UTR) of the mRNAs. The miRNAs are derived from precursor transcripts of ~70–120 nts sequences that fold to form stem-loop structures believed to be highly conserved in genome evolution. Recent research suggests that miRNAs comprise at minimum 1% of all human genes and regulate 10% or more of all human protein coding genes [12]. A number of miRNAs have been shown to play critical roles in time development, cell death, cell proliferation, fat metabolism, hematopoiesis and nervous system patterning in animals, and stress responses, and leaf and flower development in plants [13-17].

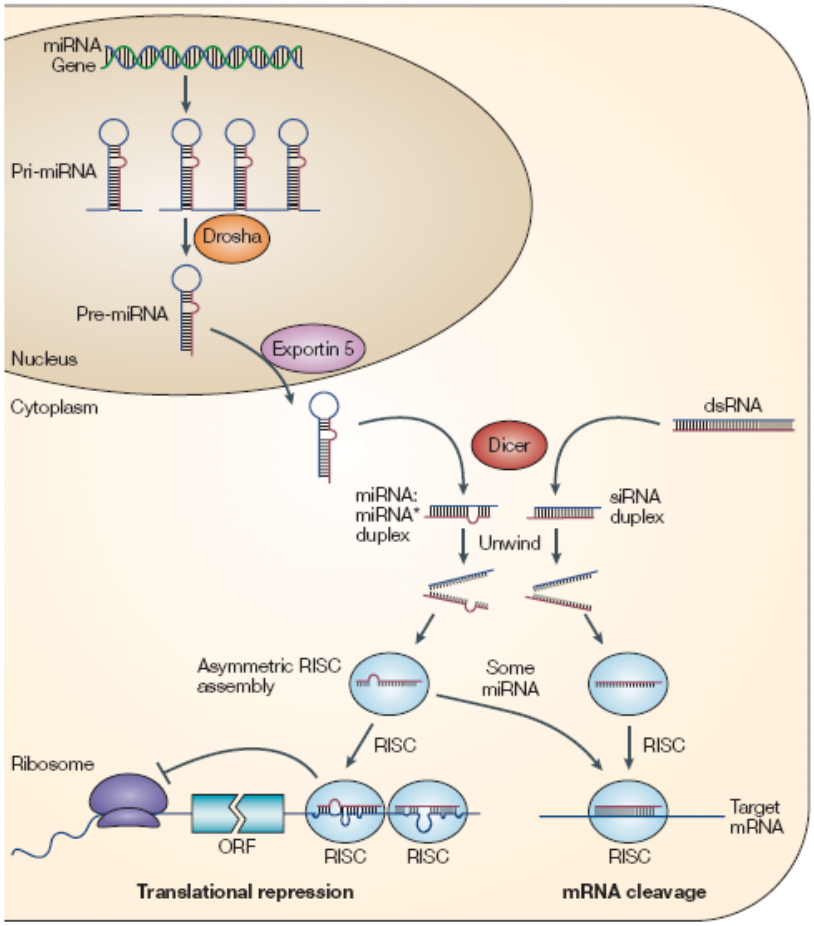


Figure 1.2 Biogenesis of miRNA [4].

1.1.3 miRNA Biogenesis

The general biogenesis of the miRNA is shown in Fig. 1.2. These miRNA genes are typically transcribed by RNA polymerase II [18]. The primary miRNA transcripts (pri-miRNAs) contain cap structures as well as the poly(A) tails, which are the unique properties of class II gene transcripts. Then, the pri-miRNAs are processed into the precursor of miRNAs (pre-miRNAs) by a protein, namely Drosha. The pre-miRNA is folded as a double-strand 'hairpin' structure which contains a short nucleotide (~17-24 nts) sequences embedded in the stem region. The pre-miRNA is exported from the nucleus to the

cytoplasm by Exportin 5. The pre-miRNA is then processed by the enzyme DICER into a dsRNA (double strand RNA) that includes the mature sequence and its partially homologous complement. This dsRNA is further processed to the mature sequence, which becomes part of the RNA-induced silencing complex (RISC).

1.1.4 Post-transcriptional Repression by miRNAs

In animals, miRNAs are imperfectly complementary to the target mRNA which usually locates in 3'-untranslated region (3'-UTR). In plants, miRNA are perfectly complementary to the target mRNA. As shown in Fig. 1.3, while miRNA/mRNA interactions with perfect complementarity tend to result in mRNA cleavage and degradation; and miRNA/mRNA interactions with imperfect complementarity tend to result in blocking ribosome processing and inhibiting mRNA translation.

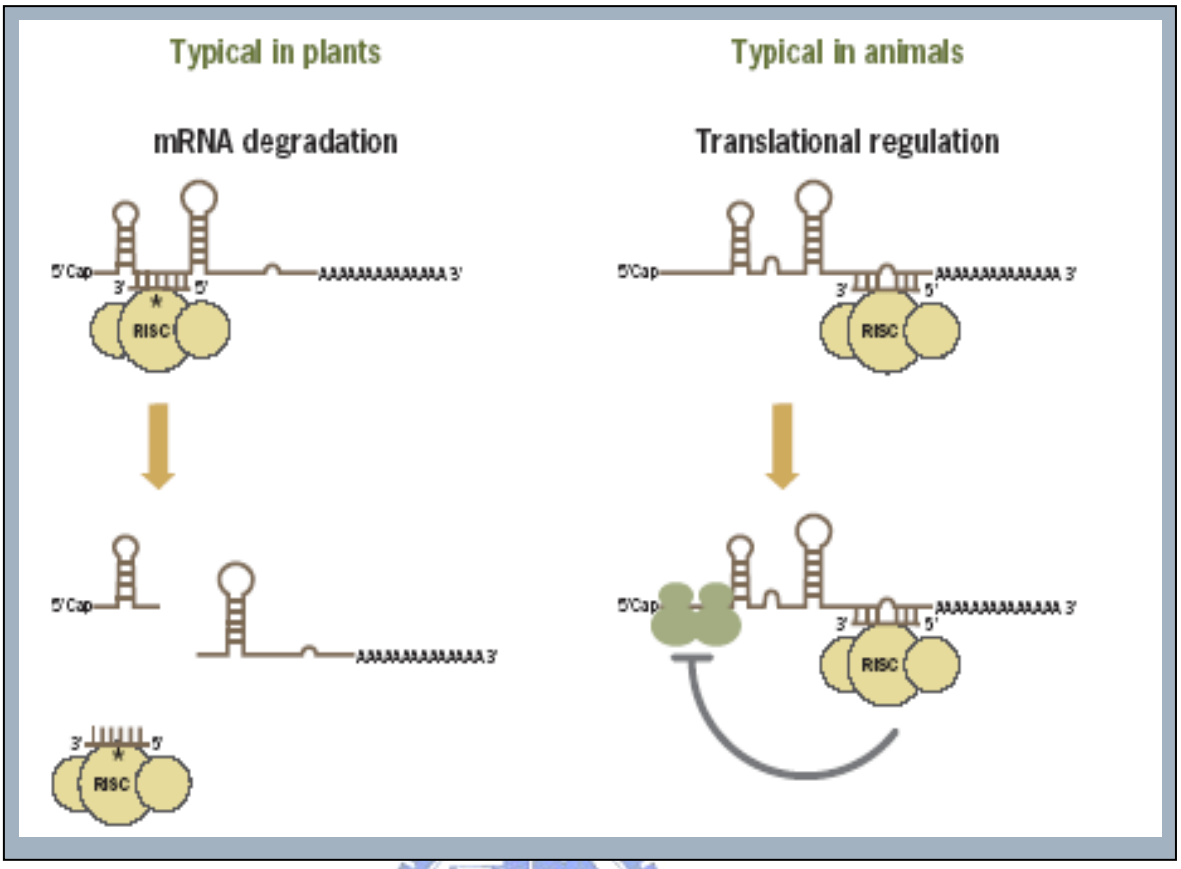


Figure 1.3 miRNA regulation function.

Since the discovery of the *lin-4* regulates the *lin-14* expression by hybridizing to the 3'-UTR of the *lin-14* gene, there have been numerous descriptions of mature miRNA are partially complementary to the 3'-UTR of a protein coding gene. A regulatory role of miRNA affecting mRNA degradation and stability (in plants) and translation (in animals) is now known [14, 19]. In addition to interfering with translation of mRNA, miRNA may conversely increase gene expression by binding with some other regulatory RNA, in effect inhibiting inhibition [20]. Previous investigation reported that some part of the miRNA region are the most important when the miRNA bound to its target. This region is nucleotides 2-8 of miRNA, called seed

region, as shown in Fig. 1.4.

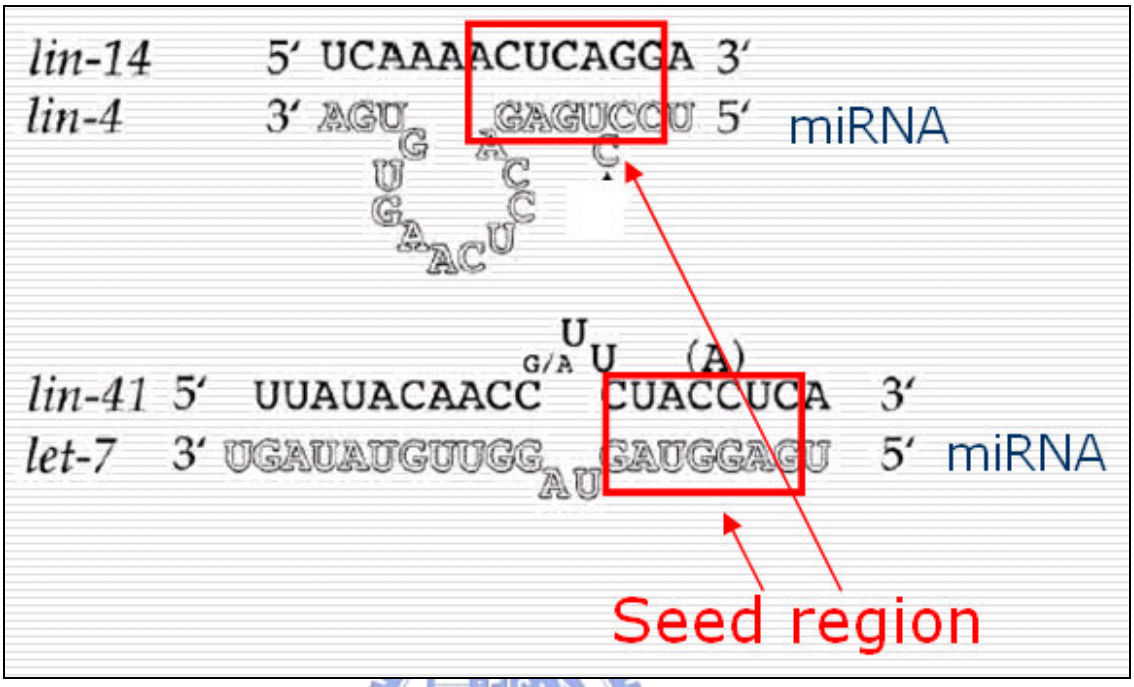


Figure 1.4 The seed region of the miRNA.

1.2 Motivation

MicroRNAs (miRNAs) bind to specific sites of target mRNAs resulting in cleavage of the message or inhibit translation. The specific function of most miRNAs is unknown. Prediction of miRNA targets provides an alternative approach to assign biological functions. Many computational tools have been previously developed for predicting miRNA target sites. However, these tools, such as miRanda [12], TargetScan [21] and RNAhybrid [22], do not perform efficiently enough to become a web server that can facilitate the identification of miRNA targets for biologists. Owing to provide the convenient analysis for biologists via a web server, more efficient miRNA target prediction tools are crucial. This work is to develop an efficient microRNA target prediction web server to provide efficient and convenient analysis for the investigators who are interested in the regulation of miRNAs.

1.3 The Specific Aim

In this work, we propose a fast and more efficient approach to predict the miRNA target sites and provide the information about the secondary structure of mRNA containing miRNA target site.

The primary contribution of this work is to successfully develop an efficient microRNA target prediction web server for identifying miRNA target sites in the conserved sequences of mammalian genomes. User-friendly

interface are designed for displaying the information of the resulted miRNA targets. Moreover, this work also provides additional information about the RNA secondary structures of the mRNA containing the miRNA target site, which can be targeted by miRNAs.

Chapter 2 Related Works

Finding regulatory mRNA targets is essential to understand the biological functions of miRNAs. Different methods were developed to predict the miRNA targets in animals or plants. In this Chapter, we introduce the difference among the miRNA target prediction methods, such as miRanda, RNAhybrid and TargetScan, that are available for the prediction of miRNA targets. This work roughly divides these methods into three categories: database, web server and software (Table 2.1).

Category	Name	Species	Reference
Database	miRNAMap	Mammalian	[23]
	miRBase	Metazoa	[24]
	TarBase	Mammalian, Nematodes, Plants, Viruses	[25]
Web Server	miRU	Plants	[26]
	MicroInspector	Arthropods, Vertebrates, Plants, Nematodes, Viruses	[27]
Software	miRanda	Human, Flies	[12]
	RNAhybrid	Flies	[22]
	TargetScan	Vertebrates	[21]

Table 2.1 The types of related miRNA target prediction.

2.1 miRNA target Databases

2.1.1 miRNAMap

In previously research of our group, miRNAMap [23] establishes genomic maps for microRNA precursors and their mapping to targets in vertebrate genomes. In this system, this work collects the known miRNAs, which are experimentally confirmed, from the miRBase [28] and predict putative miRNA precursors by RNAz which is based on genome-wide mapping of conserved RNA secondary structures from UCSC clustering [29]. This work also predicts the mature miRNAs by an algorithm which is based on MDD (maximal dependence decomposition), all of the putative and known mature miRNAs have been detected their targets by miRanda [12] results.

The prediction approach of miRNAs and miRNA targets consists of three main phases and the system flow is shown in Fig. 2.1. The three phases are the preprocessing phase, the alignment phase, and the filtering phase. In the preprocessing phase, the gene genomic sequences, EST, mRNA sequences and protein sequences, which are stored in different biological databases, are collected, converted and integrated

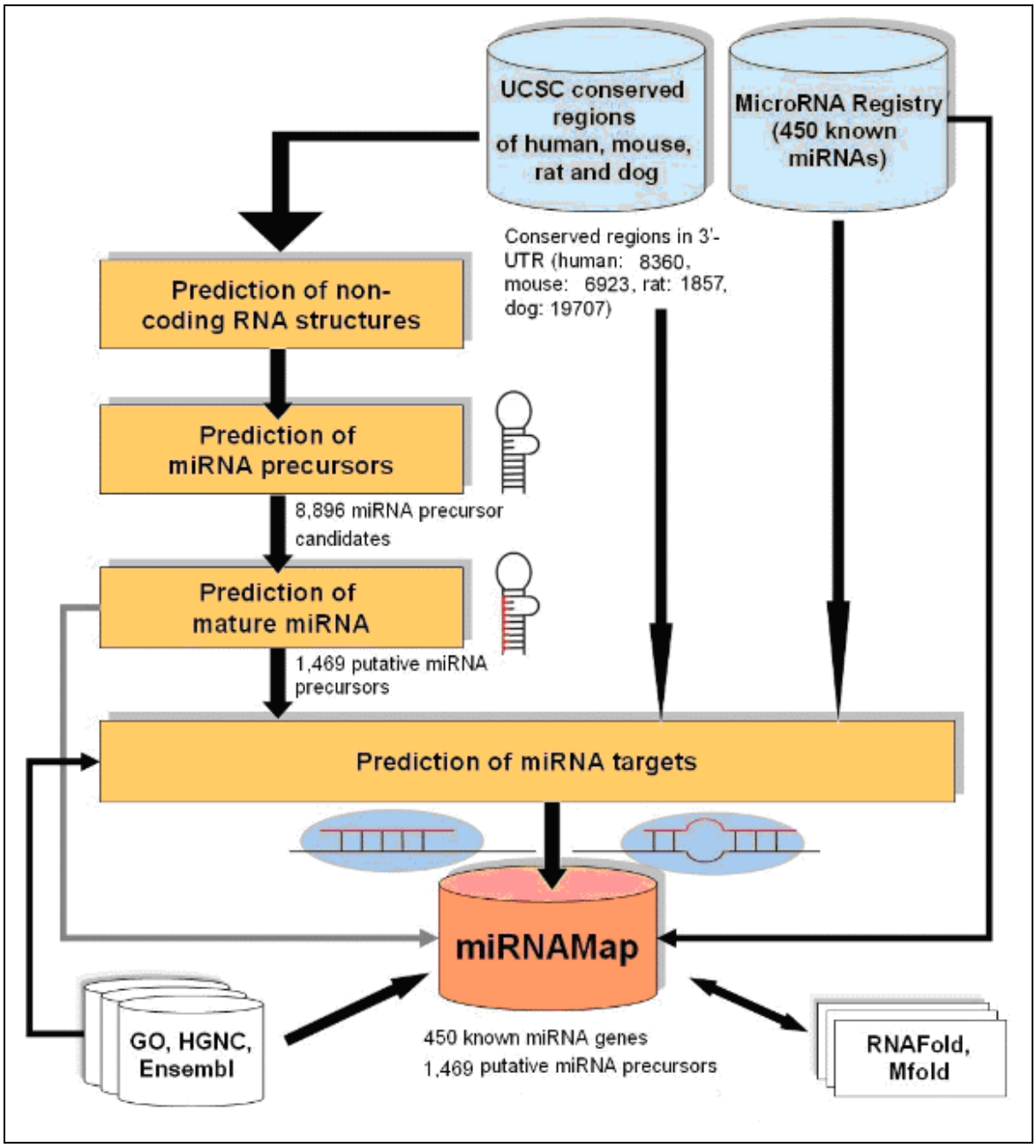


Figure 2.1 System flow of the miRNAMap

2.1.2 miRBase:Targets

As focus shifts from miRNA gene identification to functional characterization, miRBase [24] includes not only miRNA sequence data but also information about their genomic targets. The core predictions are

generated in-house using the miRanda algorithm (v3.0) [12]. The strengths of miRanda are that it is open source, scalable and incorporates robust statistical models. The provision of a P-value for each miRNA–target assignment allows the user to assess the confidence in the prediction.

2.1.3 TarBase

TarBase is a first online database for systematic collection and description of miRNA targets with experimental support. They collect the experimentally verified miRNA targets in human/mouse, fruit fly, worm, and zebrafish.

They emphasize their database will not only be useful for biologists interested in miRNA function, but also for bioinformaticians interested in using the most comprehensive set of supported targets currently available to train and test a new cohort of machine-learning methods for target prediction.

They do provide the most comprehensive of experimentally verified animal microRNA targets.

2.2 miRNA Target Prediction Web Servers

2.2.1 miRU

miRU [26] is a web-based integrated computing system. It aims at predicting plant miRNA targets. Users can input the plant mature miRNA

sequence, the system thoroughly searches for potential complementary target sites with mismatches tolerable in miRNA-target recognition. The miRU web server is available at http://bioinfo3.noble.org/miRU.htm.

Because the plant miRNA seems to hybridize almost perfectly to its cognate mRNA, it is possible to predict the plant miRNA target in the web server format.

2.2.2 MicroInspector

MicroInspector [27] is also a web tool to predict the miRNA targets online. Users could input the only one user-defined RNA sequence, which is typically an mRNA or a part of an mRNA, and then further to analyze the potential miRNA target site on this mRNA. MicroInspector [27] allows for variations in temperature, energy values and allows selection of different miRNA databases for identifying miRNA binding sites of different strength. The service can be accessed via http://www.imbb.forth.gr/microinspector.

2.3 miRNA Target Prediction Software

2.3.1 miRanda

The miRanda [12] miRNA target prediction is based on a three-phase analysis pipeline. The three phases are shown in Fig. 2.2 as follows: sequence-matching to assess first whether two sequences are complementary

and possibly bind; free energy calculation to estimate the energetics of this physical interaction; and evolutionary conservation as an informational filter.

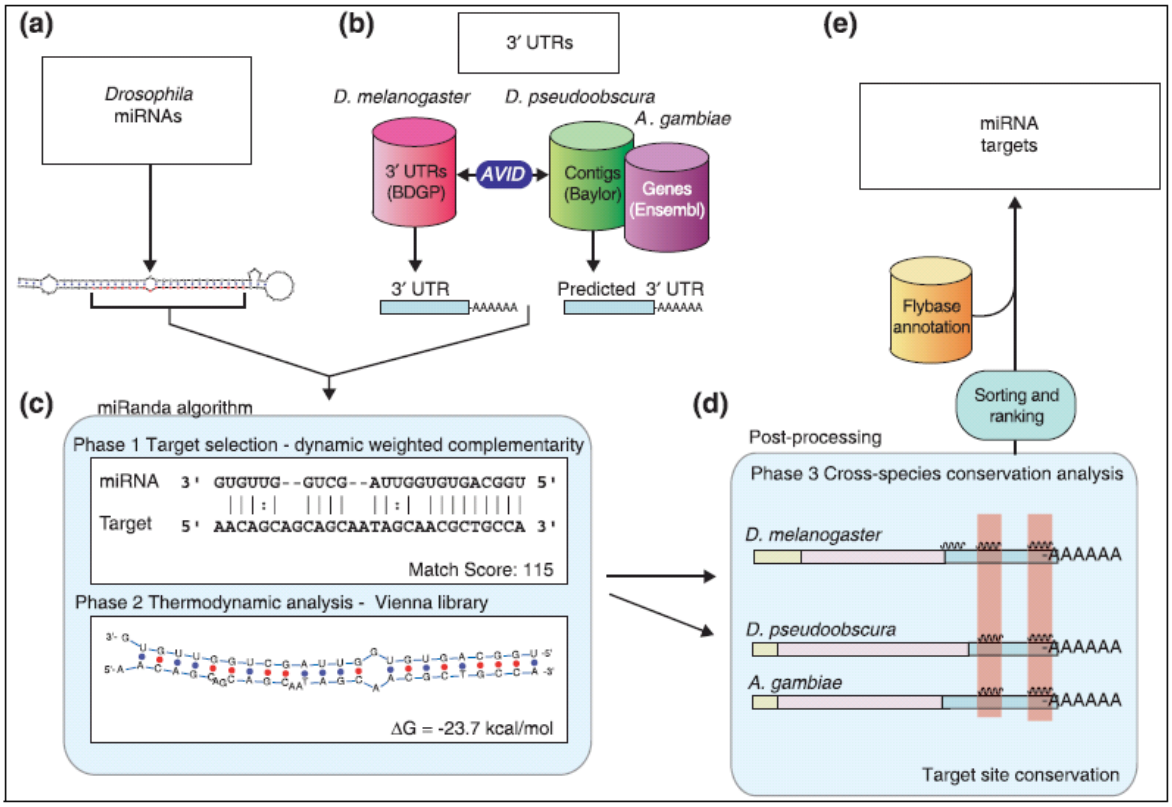


Figure 2.2 miRanda algorithm and analysis pipeline.

In 2004, the authors use miRanda to predict the human microRNA targets, and they suggest that miRNA genes, which are about 1% of all human genes, regulate protein production for 10% or more of all human genes.

2.3.2 RNAhybrid

RNAhybrid [22] predicts multiple potential binding sites of miRNAs in large target RNAs. In general, the program finds the energetically most favorable hybridization sites of a small RNA in a large RNA. Intra-molecular hybridizations, that is, base pairings between target nucleotides or between

miRNA nucleotides are not allowed. For large targets, the time complexity of the algorithm is linear in the target length, allowing many long targets to be searched in a short time. Statistical significance of predicted targets is assessed with an extreme value statistics of length normalized minimum free energies, a Poisson approximation of multiple binding sites, and the calculation of effective numbers of orthologous targets in comparative studies of multiple organisms. RNAhybrid is available at http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/.

2.3.3 TargetScan

Lewis et al [21], predict regulatory targets of mammalian microRNAs by identifying mRNAs with conserved complementarity to the seed (nucleotides 2-8) of the miRNA. They developed a algorithm called TargetScan, which combines minimum free energy of miRNA/mRNA duplex with comparative sequence analysis to predict miRNA targets conserved across multiple genomes.

Chapter 3 Materials and Method

3.1 Materials

The miRTar integrates several biological data sources and software. Table 3.1 gives the data sources integrated into miRTar. Table 3.2 gives several integrated software applied in miRTar, such miRanda, RNAhybrid and TargetScan.

Category	Data Sources	URL	Reference
Genome Sequence	Ensembl	http://www.ensembl.org/	[30]
Known miRNA Sequence	miRBase	http://microrna.sanger.ac.uk/sequences/	[24]
Conserved Sequence	UCSC Genome Browser	http://genome.ucsc.edu/	[31]
Gene Product Description	GO	http://www.geneontology.org	[32]

Table 3.1 Database comparison of materials.

Category	Name	Version	Reference
miRNA Target Prediction	miRanda	V1.9	[12]
	TargetScan	V1.0b	[21]
	RNAhybrid	V2.1	[22]
RNA Secondary Structure	mfold	V3.2	[33]
	Vienna Package	V1.6.1	[34]

Table 3.2 Tool comparison of materials.

Alignments of conserved sequences

Genome-wide alignments of vertebrates (‘multiz17way’) were

downloaded from the UCSC Genome Browser [31]. These alignments included sequences of seventeen species: human, March 2006 (hg18); chimpanzee, November 2003 (panTro1); macaque, January 2006 (rheMac2); mouse, February 2006 (mm8); rat, June 2004 (rn4); rabbit, May 2005 (oryCun1); dog, May 2005 (canFam2); cow, March 2005 (bosTau2); armadillo, May 2005 (dasNov1); elephant, May 2005 (loxAfr1); tenrec, July 2005 (echTel1); opossum, January 2006 (monDom4); chicken, February 2004 (galGal2); frog, October 2004 (xenTro1); zebrafish, May 2005 (danRer3); tetraodon, February 2004 (tetNig1); fugu, August 2002 (fr1).

Reference Species	Multiple alignment of species	Number
Human	Mouse, rat, rabbit, chimp, macaque, dog, cow, armadillo, elephant, tenrec, opossum, chicken, frog, zebrafish, tetraodon, Fugu	17
Mouse	Rat, rabbit, human, chimp, macaque, dog, cow, armadillo, elephant, tenrec, opossum, chicken, frog, zebrafish, tetraodon, Fugu	17
Dog	Human, mouse, rat	4
Chicken	Human, mouse, opossum, X. tropicalis, zebrafish, tetraodon	7
Zebrafish	Tetraodon, fugu, human, mouse	5

Table 3.3 The number of multiple alignments of assemblies to each species.

This work selected the human, mouse, dog, chicken and zebrafish to support the miRNA target prediction service. According to the species considered in this work, the different number of multiple alignments of

assemblies to each species which are listed in Table 3.3.

Selection of the most conserved regions

This work retrieves the 'Most Conserved' track generated by the PhastCons program. This track was intersected with the 3'-UTR, 5'-UTR and CDS(coding region sequence) of Ensembl genes (Build 37) separately.

Known miRNA - miRBase::Sequence

The miRBase sequence database contains sequences of all published mature miRNA sequences, together with their predicted source hairpin precursors and annotation relating to their discovery, structure and function. The database has grown rapidly in the recent years, from 506 entries representing miRNA hairpin precursors in six species (release 2.0, June 2003) to 3,518 entries in 36 species (release 8.0, February 2006).

The known miRNAs in this work are obtained from miRBase (release 8.0, February 2006) in order to support the user input of miRBase [24] accession number for retrieving the mature miRNA sequences of known miRNAs. The statistics of the known miRNA sequence is shown in Table 3.1.

	Human	Mouse	Dog	Chicken	Zebrafish
# of miRNAs	328	266	5	125	177

Table 3.4 The statistics of known miRNA.

Gene Ontology

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

The three organizing principles of GO are molecular function, biological process and cellular component. A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components. And it is easy to confuse a gene product and its molecular function, because very often these are described in exactly the same words.

Mfold

Mfold (Version 3.1.2) is a tool for predicting the secondary structure of RNA sequence using thermodynamic methods. The ‘m’ simply refers to ‘multiple’. The core algorithm predicts a minimum free energy as well as minimum free energies for foldings that must contain any particular base pair. Base-paris within this free energy increment are chosen either automatically or else by the user. Then foldings that contain the chosen base pair are

computed.

They also provided a web system for the prediction of the secondary structure of single stranded nucleic acids. The objective of this web server is to provide easy access to RNA and DNA folding and hybridization software to the scientific community at large.

Structure annotation has been described by Zuker and Jacobson. Bases in plotted structures may be annotated by ‘p-num’, which represent the number of ways that a base may pair in foldings from the minimum energy. Low values indicate ‘well-defined’ base. Values of 0 or 1 indicate that a base is always single stranded or always paired to a unique partner. The number of times that a base is single stranded in the computed foldings is called its ‘ss-count’ number, and structure plots may also be annotated using these numbers.

3.2 System Flow

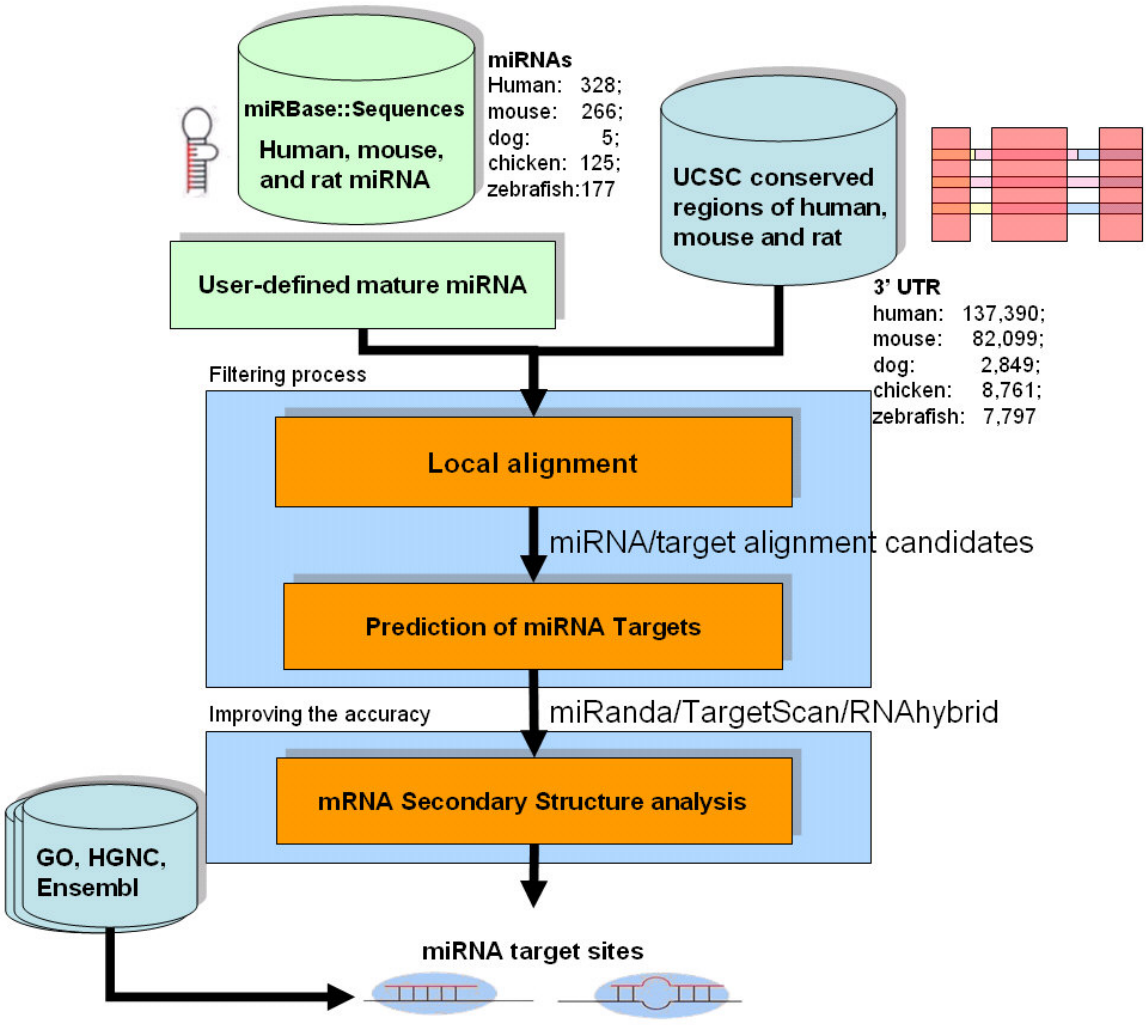


Figure 3.1 The flow of the miRTar system.

Fig. 3.1 presents the data flow of the proposed web server for miRNA target prediction. Both miRBase [24] and miRNAMap [23] are comprehensive information repository for the miRNAs and their targets. The known miRNAs in this work are obtained from miRBase (release 8.0, February 2006) in order to support the user input of miRBase accession

number for retrieving the mature miRNA sequences of known miRNAs. The conserved sequences of human, mouse and rat genomes were extracted from UCSC Genome Browser [31]. Moreover, the annotation of 5'-UTR, 3'-UTR and open reading frames (ORFs) for all transcripts of all genes are obtained from the Ensemble database [30]. The conserved regions in 3'-UTR of the genes, which are at least 200 bp long and have sequence identities greater than 60% were extracted. Consequently, there are 137390, 82099, 2849, 8761 and 7797 3'-UTR conserved regions extracted from human, mouse, dog, chicken and zebrafish genomes, respectively.

Before using miRanda/RNAhybrid/TargetScan to identify the miRNA targets against the prepared conserved sequences, a filtering strategy implemented as a sequence local alignment program based on dynamic programming is employed. In order to maximize the efficiency of the proposed web service, the simple sum-of-pair scoring function (SP scoring function) is applied for measuring the quality of the alignment. The scores of 6, 4 and 2 are assigned for G:C, A:T and G:U pairs, respectively; whereas penalties of -3 and -5 are assigned for mismatched pairs and a gap penalty, respectively. The filtering process can filter all the short fragments in a sequence database that their scores of alignment to a mature miRNA sequence exceed the score cutoff. The resulting fragments are the candidates for miRNA targets, which are then used as the search database when applying miRanda, RNAhybrid and TargetScan for predicting miRNA targets.

Following the filtering process, miRanda, RNAhybrid and TargetScan were integrated into the proposed miRTar to identify the miRNA targets among the candidates of miRNA targets generated in the filtering phase. The Minimum Free Energy (MFE) of the miRNA–target duplex was determined in predicting the miRNA target sites. The lower MFE values of the miRNAs and the target sites reveal the energetically more probable hybridizations between the miRNAs and the target genes. Additionally, the predictive parameters including miRanda MFE and miRanda score were adjusted for the miRNA target prediction by comparing the predictive results to known miRNA/targets data during our previous work [23]. The MFE threshold of the miRNA and target duplex was suggested as -16 kcal/mol and the miRanda score was specified as 160. The miRNA targets corresponding to a miRNA whose MFEs are smaller than -16 kcal/mol and the score exceeds 160 are thus identified.

3.3 Predict the Structure of the Target mRNA

Previous report [35] hypothesized that single-stranded miRNAs can only search stretches of free mRNA for potential target sites. If a stretch of RNA is unbound in one state and bound in the other, the probability of binding is relatively high. On the other hand, if the mRNA is folded so that the site of interest is based paired with another part of the mRNA, then the energy difference between the two states is smaller. Of course, there are proteins

wrapping the miRNA that could potentially play a role in recognition. However, there is no evidence that the relevant proteins recognize either sequence or structure of the mRNA targets.

This work proposes a concept to describe the free energy difference between the miRNA hybridizes to its target site and miRNA unhybridizes to its target site. In other word, it means that miRNA will contribute additional free energy to stabilize the target mRNA. Moreover, the detected miRNA targets are presented in the web server. In particular, to provide additional information about the hybridization of miRNAs and miRNA targets, the five measures, E0, E1, E2, E3 and E4, are defined and used to examine the change of MFE in five different situations when a miRNA hybridize to its target (Fig. 3.2). The E0 is the MFE of the miRNA target site extending in two directions. The E1 is the MFE of the hybridization of a miRNA and a miRNA target site. The E3 is the MFE of the RNA secondary structure generated from the surrounding sequences of the miRNA target site with the replacement of the miRNA target site by a series of “N”, which indicates none nucleotides (pseudo nucleotides). The E2 is the MFE of the secondary structure generated from the surrounding sequences of miRNA target sites with the insertion of ‘N’ nucleotides. The E4 is the MFE of the secondary structure to simulate the hybridizing structure of a miRNA and a miRNA target site combined to the RNA secondary structure of the sequence surrounding to the miRNA target sites. Figure 3.2 shows the calculations for E0, E1, E2, E3 and E4. This work

hypothesize that the value C = (|E4| - |E2|) indicates the contribution of energy when a miRNA binds to it target site. The higher C value, the more probable the miRNA target site of a miRNA.

Furthermore, especially the diagram of the RNA secondary structure of the mRNA precursor is folded using RNAfold [34] and is generated graphically by mfold [33]. It will predict one best secondary structure and some suboptimal secondary structures and this work used the best one to show in web system.

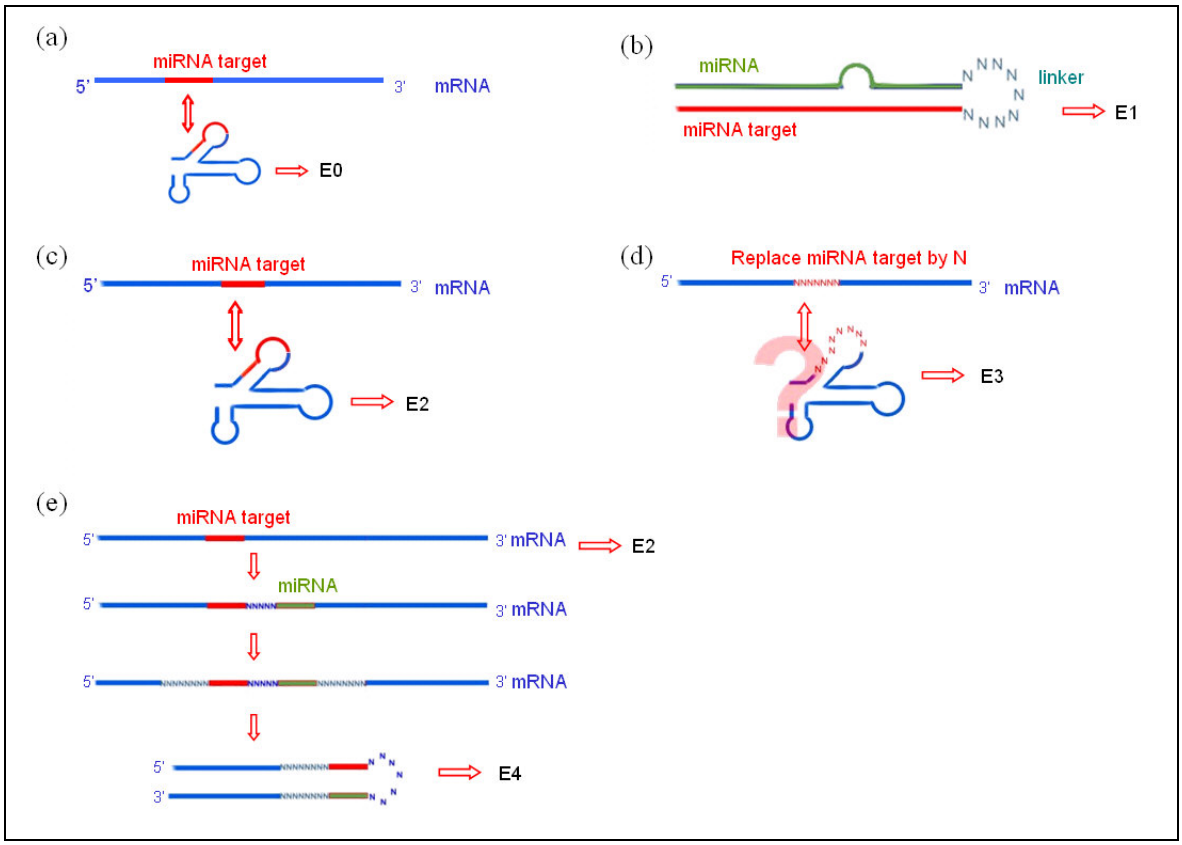


Figure 3.2 Definition for the measurement of E0, E1, E2, E3 and E4.

3.4 Implementation Environment

The web server is currently implemented on a PC server with dual-processor 2.8 GHz Intel Xeon and 2 GB RAM memory. The operation system is Red Hat Linux.

Chapter 4 Results

4.1 Performance Evaluation

4.1.1 Comparing miRTar – filtering process with miRanda

To compare the search efficiency of miRanda and miRTar, 30 human miRNAs were experimented in both programs to predict their targets in human conserved 3'-UTR sequences (Fig. 4.1). The spent time of the two program are calculated with different MFE cutoff (-16– -20, interval -1) and different miRanda scores cutoff (160–200, interval 10) (Fig. 4.2). During this experiment, the size of the conserved 3'-UTR sequences is roughly 24 Mbps. The miRanda required approximately 105 seconds to accomplish the overall analysis; whereas the miRTar only required about 23 seconds. Both miRanda and miRTar performed stably by setting different predictive parameters, i.e., MFE cutoff and miRanda score cutoff.

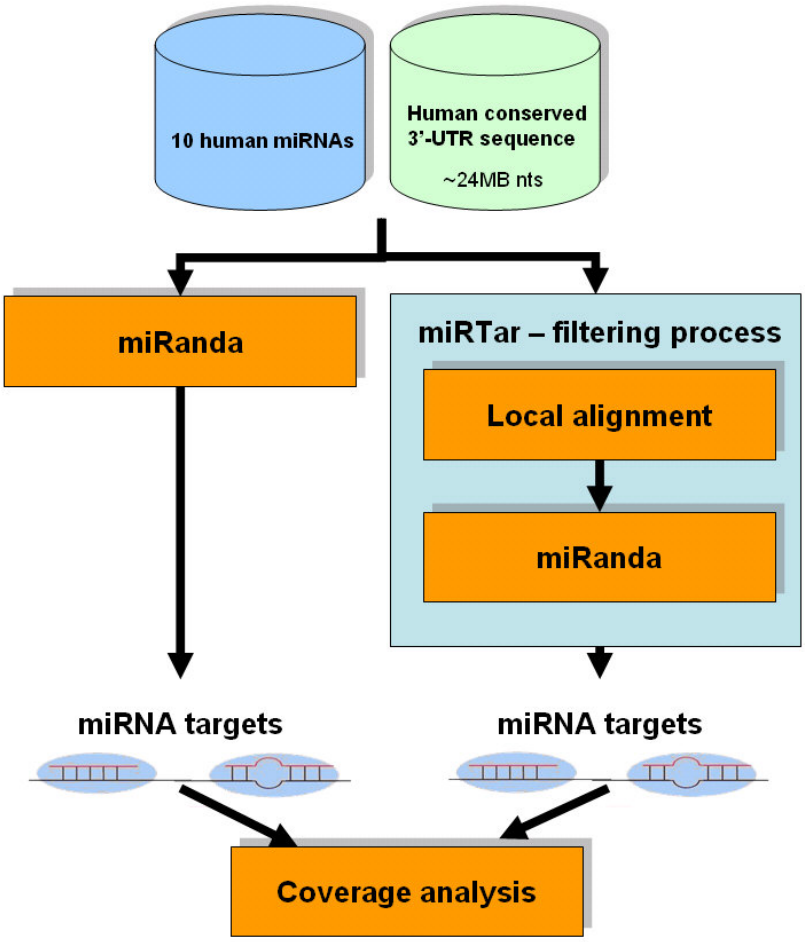


Figure 4.1 The flow of performance evaluation for miRTar – filtering process and miRanda.

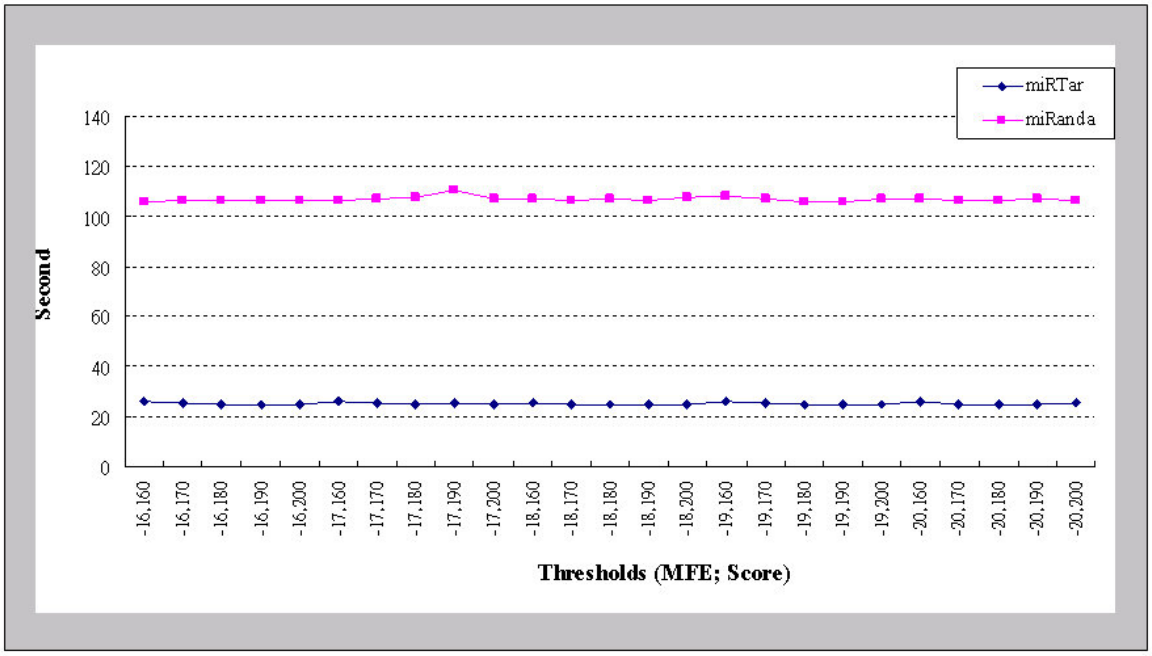


Figure 4.2 The efficiency comparison between miRTar and miRanda.

Moreover, the coverage analysis of the predicted results using miRTar and miRanda were performed for assessing the level of data loss during the filtering process proposed. Table 4.1 gives the analytical results of the coverage analysis. For example, 287 miRNA targets of hsa-let-7a were identified by miRanda and 273 miRNA targets were identified by our miRTar. Comparing the predicted results by the two programs, 260 miRNA targets are identified by both programs and the coverage ratio is 90.59%. Figure 4.3 summarizes the coverage of the predicted miRNA targets by the two programs. Thus, the average coverage of the predicted results between miRTar and miRanda is about 97%.

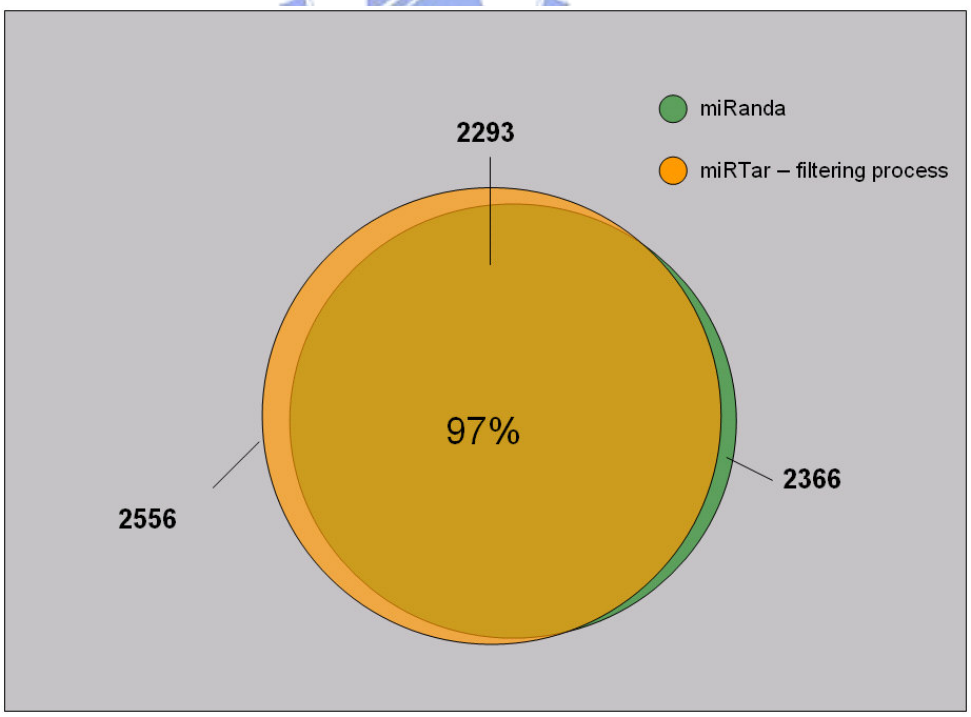


Figure 4.3 Coverage analysis of the prediction results between miRTar – filtering process and miRanda.

miRBase miRNAs	No. of targets predicted by miRanda (A)	No. of targets predicted by miRTar – filtering process (B)	No. of targets predicted by both miRTar – filtering process and miRanda (C)	Coverage ($\frac{C}{A}$) (%)
hsa-miR-130b	106	104	104	98.11%
hsa-miR-184	338	336	311	92.01%
hsa-miR-200b	17	17	17	100.00%
hsa-miR-205	186	190	186	100.00%
hsa-miR-222	202	226	195	96.53%
hsa-miR-345	539	557	533	98.89%
hsa-miR-423	760	903	735	96.71%
hsa-miR-511	41	45	41	100.00%
hsa-miR-526b	176	177	170	96.59%
hsa-miR-9*	1	1	1	100.00%
Total	2,360	2,556	2,293	Average: 96.91%

Table 4.1 Prediction coverage between miRTar – filtering process and miRanda.

4.1.2 Comparing miRTar – filtering process with TargetScan

To compare the efficiency between TargetScan and miRTar, 10 human miRNAs were selected and experimented in both programs to predict their targets in human ~24 MB conserved 3'-UTR sequences (Fig. 4.4). The TargetScan required approximately 22 seconds accomplishing the overall analysis. It is notable that we applied the TargetScan 1.0 beta version to perform this evaluation. TargetScan searches the "seed matches" on target sequence and optimizes basepairing of the remaining 3' portion of the miRNA to the 35 bases of the target sequence. It also computes which this target site is statistically significance or not. But the beta version of TargetScan doesn't support to assign a Z-score to each target site. We examined the prediction

results of TargetScan, there were something strange that TargetScan found the miRNA/mRNA with the large internal loop which were not the possible miRNA target site according to known miRNA/mRNA duplexes criteria.

Therefore, the average coverage of predicted results between miRTar and TargetScan is 69% (Fig. 4.5) when the total running time of two programs is about 22 seconds. The phenomenon described before maybe due to the beta version of TargetScan.

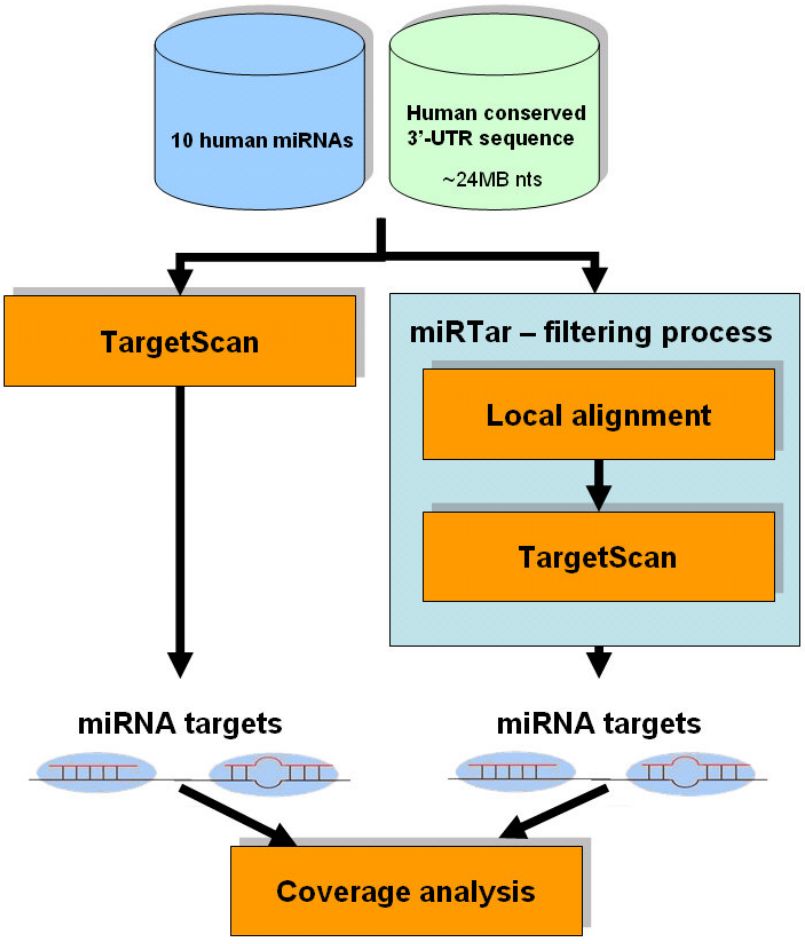


Figure 4.4 The flow of performance evaluation for miRTar –filtering process and TargetScan.

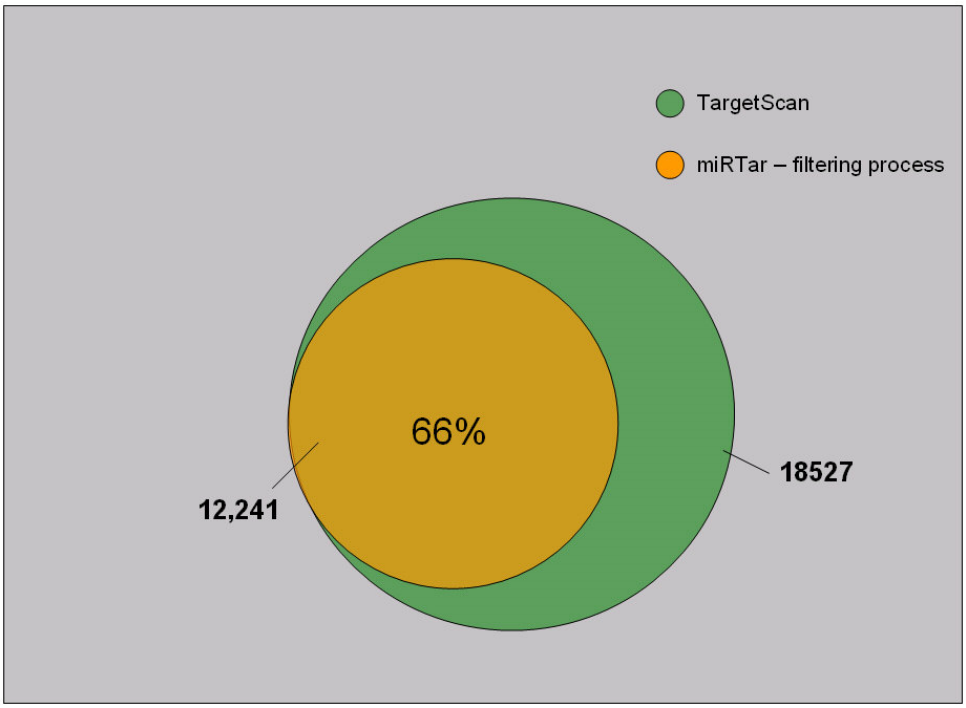


Figure 4.5 Coverage analysis of the prediction results between miRTar – filtering process and TargetScan.

miRBase miRNAs	No. of targets predicted by TargetScan (A)	No. of targets predicted by miRTar – filtering process (B)	No. of targets predicted by both miRTar – filtering process and TargetScan (C)	Coverage ($\frac{C}{A}$) (%)	
hsa-miR-130b	2,381	1,957	1,957		82.19%
hsa-miR-184	491	456	456		92.87%
hsa-miR-200b	3,057	1,242	1,242		40.63%
hsa-miR-205	1,962	1,408	1,408		71.76%
hsa-miR-222	1,289	1,029	1,029		79.83%
hsa-miR-345	1,463	1,226	1,226		83.80%
hsa-miR-423	347	307	307		88.47%
hsa-miR-511	3,311	2,288	2,288		69.10%
hsa-miR-526b	1,721	924	924		53.69%
hsa-miR-9*	2,505	1,404	1,404		56.05%
Total	18,527	12,241	12,241	Average:	66.07%

Table 4.2 Prediction coverage between miRTar – filtering process and TargetScan.

4.1.3 Comparing miRTar – filtering process with RNAhybrid

To compare the efficiency between RNAhybrid and miRTar – filtering process, 10 human miRNAs were selected and experimented in both programs to predict their targets in human ~24 MB conserved 3'-UTR sequences (Fig. 4.6).

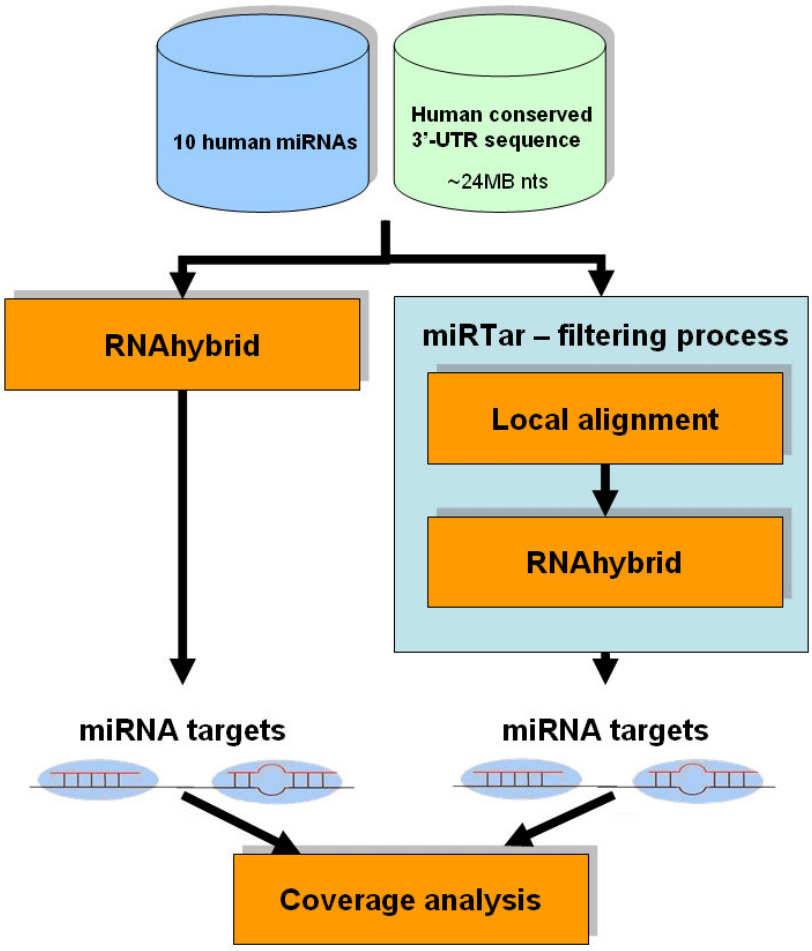


Figure 4.6 The flow of performance evaluation for miRTar – filtering process and RNAhybrid.

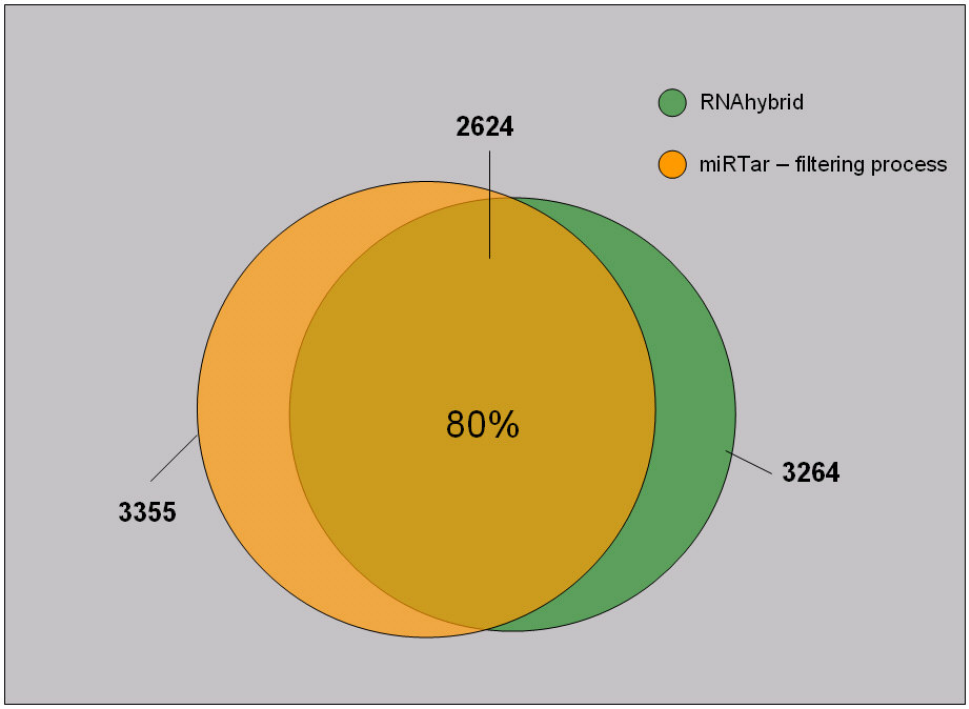


Figure 4.7 Coverage analysis of the prediction results between miRTar – filtering process and RNAhybrid.

miRBase miRNAs	No. of targets predicted by RNAhybrid (A)	No. of targets predicted by miRTar-filtering process (B)	No. of targets predicted by both miRTar – filtering process and RNAhybrid (C)	Coverage ($\frac{C}{A}$) (%)
hsa-miR-130b	37	38	25	67.57%
hsa-miR-184	122	135	100	81.97%
hsa-miR-200b	5	4	2	40.00%
hsa-miR-205	320	312	180	56.25%
hsa-miR-222	320	335	192	60.00%
hsa-miR-345	336	377	223	66.37%
hsa-miR-423	1,320	1,403	1,195	90.53%
hsa-miR-511	601	539	534	88.85%
hsa-miR-526b	202	211	172	85.15%
hsa-miR-9*	1	1	1	100.00%
Total	3,264	3,355	2,624	Average: 80.39%

Table 4.3 Prediction coverage between miRTar – filtering process and RNAhybrid.

4.2 Web Interfaces

4.2.1 Overview of Web Interface

The miRTar tool is a web-based tool for predicting miRNA targets in the conserved mRNA sequences of the genes that are potentially regulated by a user-submitted miRNA sequence. Fig. 4.7 presents the miRTar web interface for user input. Users simply input miRBase accession number for retrieving a known miRNA or a small RNA sequence user-specified as a mature miRNA sequence that is not collected in miRBase. Users then select the search regions of mRNA, such as 3’-UTRs, 5’-UTRs and the coding regions, for the sequences that possibly contain the miRNA targets. Owing to mammalian miRNAs were investigated to regulate gene expression by imperfectly base-pairing to the 3’-UTR of target mRNAs and suppressing protein synthesis [14], the conserved sequences of 3’-UTRs are chosen as the default miRNA target regions. Additionally, a recent study [36] indicated that let-7 maintains functionality even when its binding sites located into the coding regions or 5’-UTRs and suggested that the miRNA is likely complementary to its target site in the 5’-UTRs or coding regions and mediates inhibition of protein synthesis. Additionally, the proposed miRTar allows users to adjust the predictive parameters, including the hybridization temperature, MFE cutoff, and miRanda score cutoff.

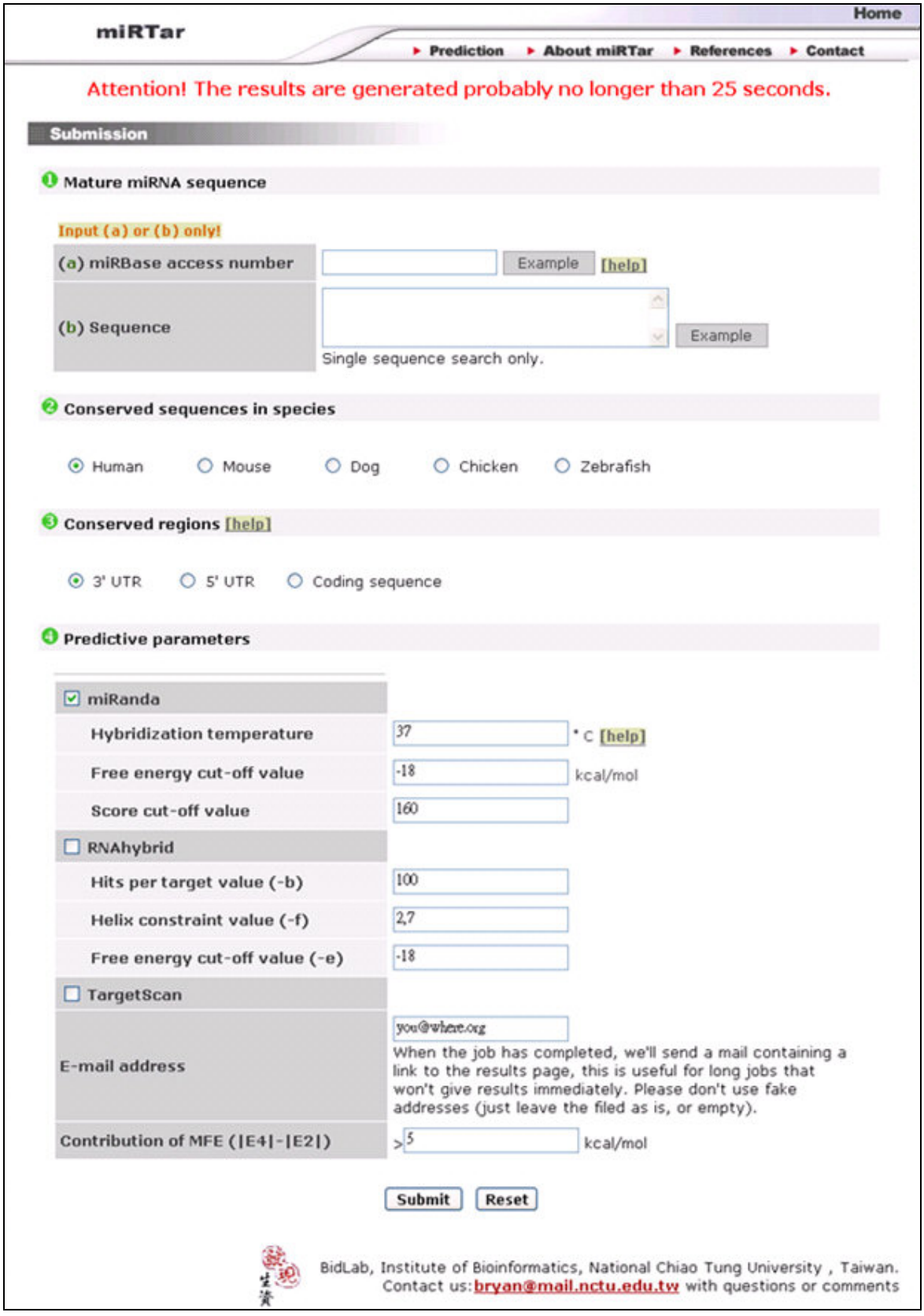


Figure 4.8 The user input interface of miRTar.

4.2.2 Immediate Execution versus Batch Jobs

Predicting miRNA targets using only one predicting algorithm may be done while the user waits. This is the default. At this time, predicting results for immediate jobs are erased 48 h after they are submitted. When more than two predicting algorithm, the batch option must be selected from the appropriate option button. The user should enter a valid email address in this case, although email addresses are always welcome since they identify users.

4.3 Case Studies

4.3.1 Hsa-let-7e Regulates The SMC1L1

This work uses an miRNA of hsa-let-7e, which is known to interact with the target site located in the 3'-UTR of SMC1L1 [37] to demonstrate that miRTar is capable of determining that hsa-let-7e can target to SMC1L1. Here, all the predictive parameters are set to the default values. Fig. 4.8 presents the principal prediction results for hsa-let-7e. The list of results is ranked by miRanda score and MFE. The "MORE INFO" button provides additional information, as shown in Fig. 4.9, including the hybridizing structure of miRNA/mRNA interaction and the RNA secondary structure generated by the miRNA target site and its surrounding sequences.

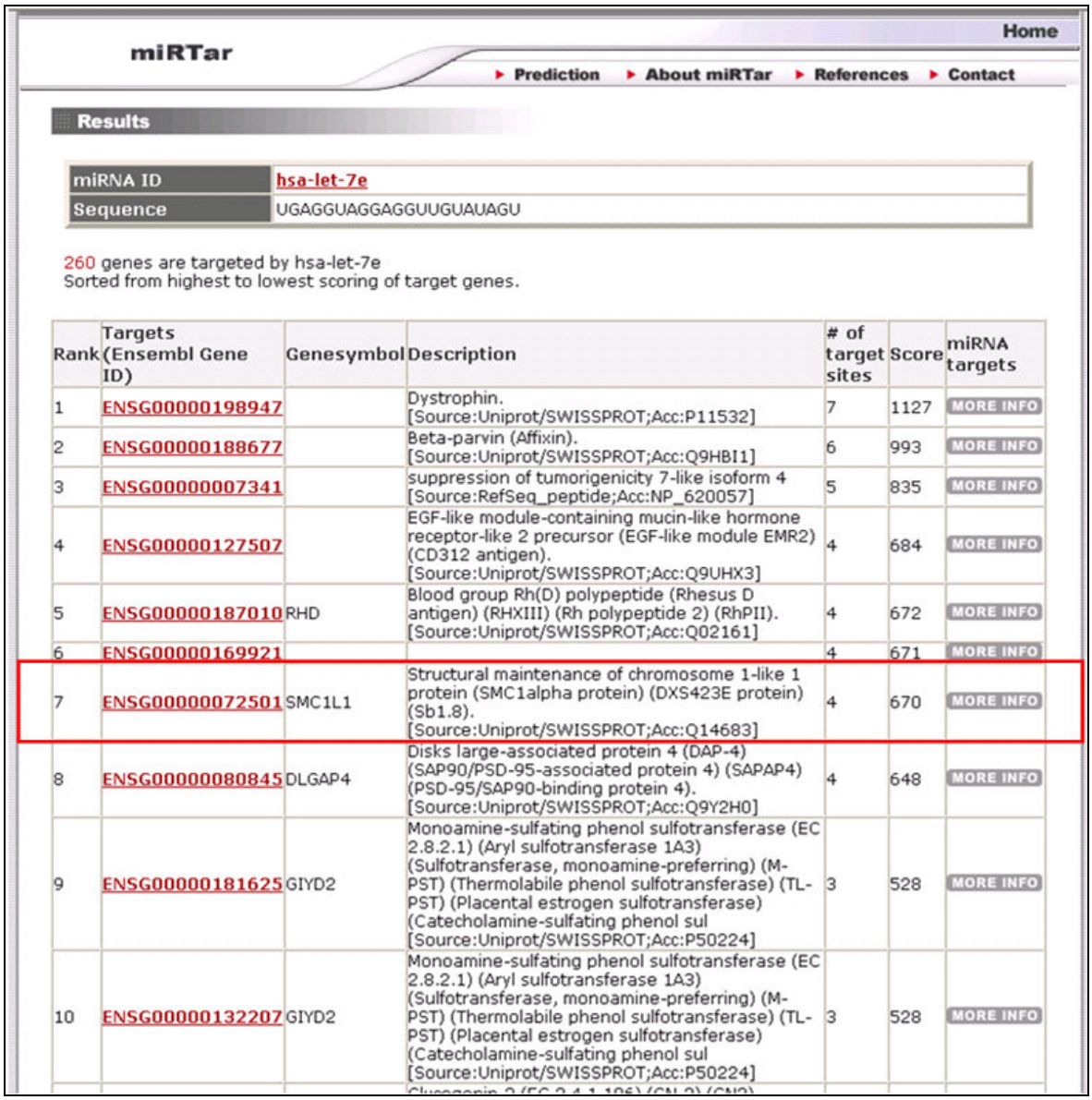


Figure 4.9 Hsa-let-7e target gene list.

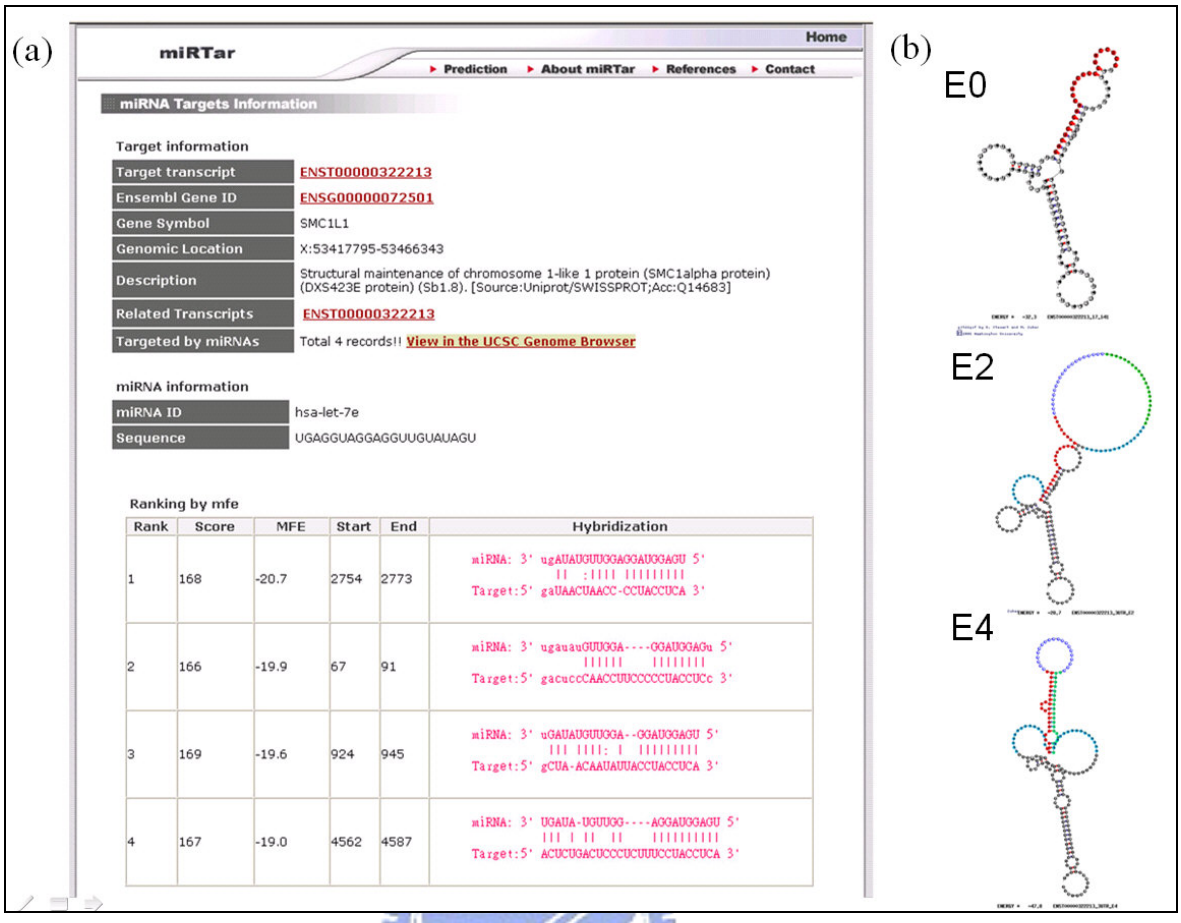
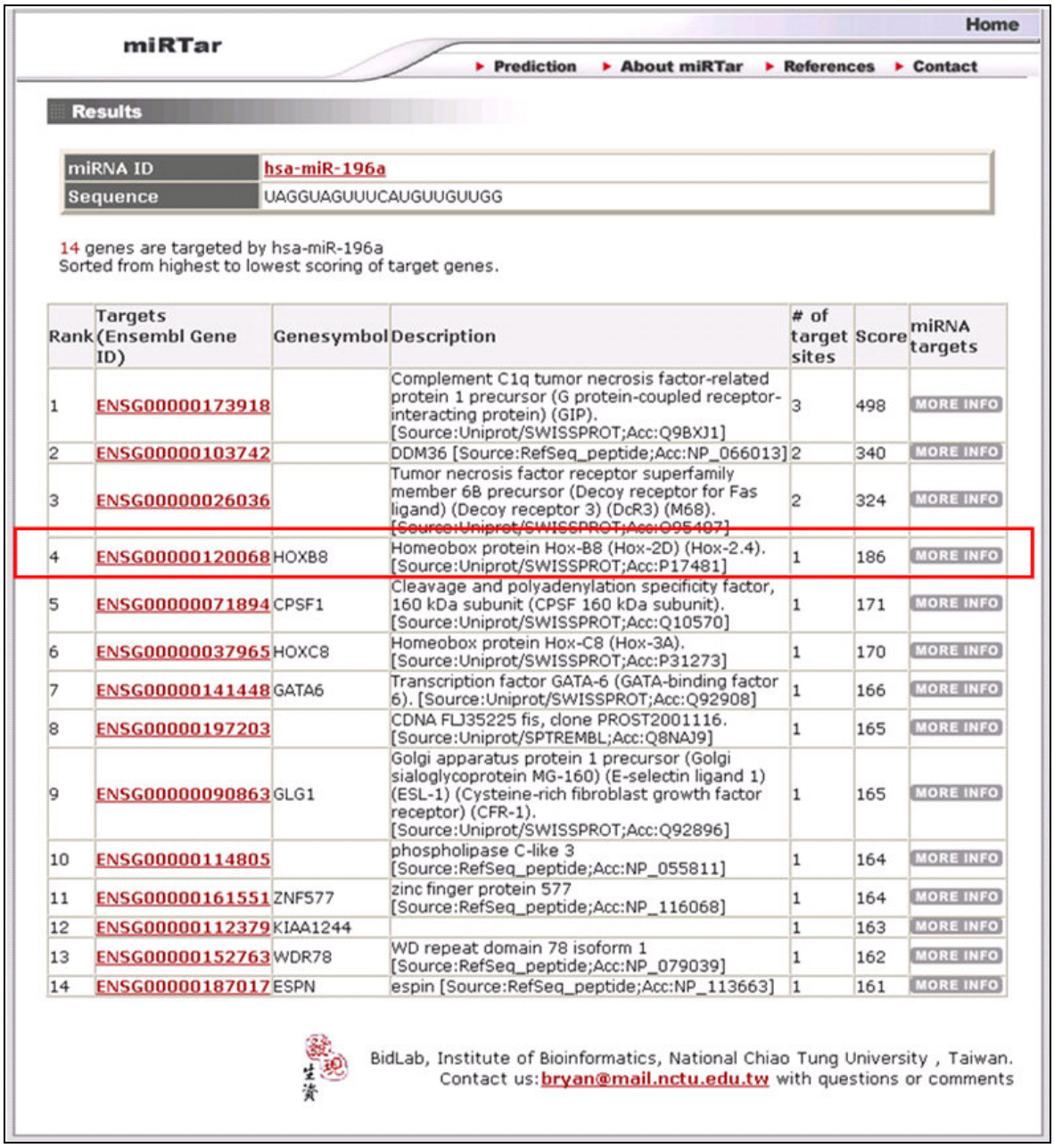


Figure 4.10 The more information of hsa-let-7e/SMC1L1.

4.3.2 Hsa-miR-196a Down-regulates HOXB8

Another example was selected to show that this work can use miRTar to predict the known miRNA/mRNA relationship. This gene is a member of the Antp homeobox family and encodes a nuclear protein with a homeobox DNA-binding domain. miR-196 acts upstream of Hoxb8 and Sonic hedgehog (Shh) in vivo in the context of limb development, thereby identifying a previously observed but uncharacterized inhibitory activity that operates specifically in the hindlimb [38]. Fig. 4.10 shows the target gene list of has-miR-196a. Detailed information is shown in Fig. 4.11.



miRTar

Home | Prediction | About miRTar | References | Contact

Results

miRNA ID	hsa-miR-196a
Sequence	UAGGUAGUUUCAUGUUGUUGG

14 genes are targeted by hsa-miR-196a
Sorted from highest to lowest scoring of target genes.

Rank	Targets (Ensembl Gene ID)	Genesymbol	Description	# of target sites	Score	miRNA targets
1	ENSG00000173918		Complement C1q tumor necrosis factor-related protein 1 precursor (G protein-coupled receptor-interacting protein) (GIP). [Source:Uniprot/SWISSPROT;Acc:Q9BXJ1]	3	498	MORE INFO
2	ENSG00000103742		DDM36 [Source:RefSeq_peptide;Acc:NP_066013]	2	340	MORE INFO
3	ENSG00000026036		Tumor necrosis factor receptor superfamily member 6B precursor (Decoy receptor for Fas ligand) (Decoy receptor 3) (DcR3) (M68). [Source:Uniprot/SWISSPROT;Acc:O95407]	2	324	MORE INFO
4	ENSG00000120068	HOXB8	Homeobox protein Hox-B8 (Hox-2D) (Hox-2.4). [Source:Uniprot/SWISSPROT;Acc:P17481]	1	186	MORE INFO
5	ENSG00000071894	CPSF1	Cleavage and polyadenylation specificity factor, 160 kDa subunit (CPSF 160 kDa subunit). [Source:Uniprot/SWISSPROT;Acc:Q10570]	1	171	MORE INFO
6	ENSG00000037965	HOXC8	Homeobox protein Hox-C8 (Hox-3A). [Source:Uniprot/SWISSPROT;Acc:P31273]	1	170	MORE INFO
7	ENSG00000141448	GATA6	Transcription factor GATA-6 (GATA-binding factor 6). [Source:Uniprot/SWISSPROT;Acc:Q92908]	1	166	MORE INFO
8	ENSG00000197203		CDNA FLJ35225 fis, clone PROST2001116. [Source:Uniprot/SPTREMBL;Acc:Q8NAJ9]	1	165	MORE INFO
9	ENSG00000090863	GLG1	Golgi apparatus protein 1 precursor (Golgi sialoglycoprotein MG-160) (E-selectin ligand 1) (ESL-1) (Cysteine-rich fibroblast growth factor receptor) (CFR-1). [Source:Uniprot/SWISSPROT;Acc:Q92896]	1	165	MORE INFO
10	ENSG00000114805		phospholipase C-like 3 [Source:RefSeq_peptide;Acc:NP_055811]	1	164	MORE INFO
11	ENSG00000161551	ZNF577	zinc finger protein 577 [Source:RefSeq_peptide;Acc:NP_116068]	1	164	MORE INFO
12	ENSG00000112379	KIAA1244		1	163	MORE INFO
13	ENSG00000152763	WDR78	WD repeat domain 78 isoform 1 [Source:RefSeq_peptide;Acc:NP_079039]	1	162	MORE INFO
14	ENSG00000187017	ESPN	espin [Source:RefSeq_peptide;Acc:NP_113663]	1	161	MORE INFO

BidLab, Institute of Bioinformatics, National Chiao Tung University , Taiwan.
Contact us: bryan@mail.nctu.edu.tw with questions or comments

Figure 4.11 Hsa-miR-196a target gene list.

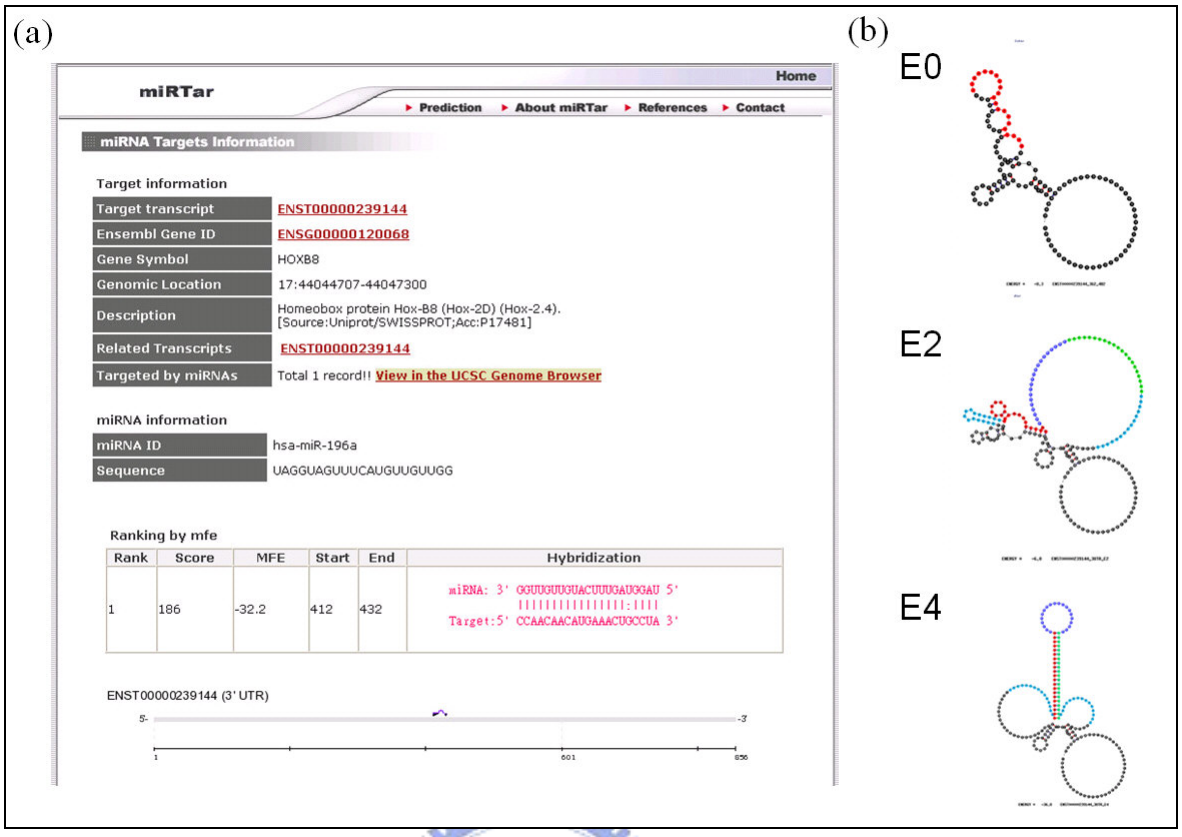


Figure 4.12 The more information of hsa-miR-196a/HOXB8.

4.4 Summary of Results

After the case studies presented above, all the demonstrations show that that our system can discovery the known miRNA/target in a more efficient way.

Chapter 5 Discussions

5.1 Limitations of miRTar

Our system has a few limitations of input data size and prediction algorithm. Basically, our system is an online web server, so there are some congenital restrictions. Limited by the computational power, the system cannot accept large datasets include multiple user-defined miRNA sequences because the miRTar requires much executing time. This work limits the user to input only one miRNA sequence every time, so it can control the waiting time in about 25 seconds. Another limitation is that this work applies the previous well-known miRNA target prediction tools to predict the possible miRNA targets, so the prediction accuracy of the miRTar is limited by those miRNA target prediction tools. To overcome this limitation, this work overlaps the miRNA targets predicted by different tools.

5.2 Defects of Adding Artificial Linker

In theory, the RNA secondary structure prediction tool will treat the Ns like the unpaired nucleotides. But if the linker Ns are add in the stem region, the secondary structure is slightly different from original one. Because the Ns will disrupt the continuous paired nucleotides and the mfe calculation will

slightly different as shown in below. The E2 (-28.7 kcal/mol) shown in Fig. 5.2 is slightly larger than E0 (-32.3 kcal/mol) shown in Fig. 5.3. In other word, the linker destabilizes the RNA secondary structure.

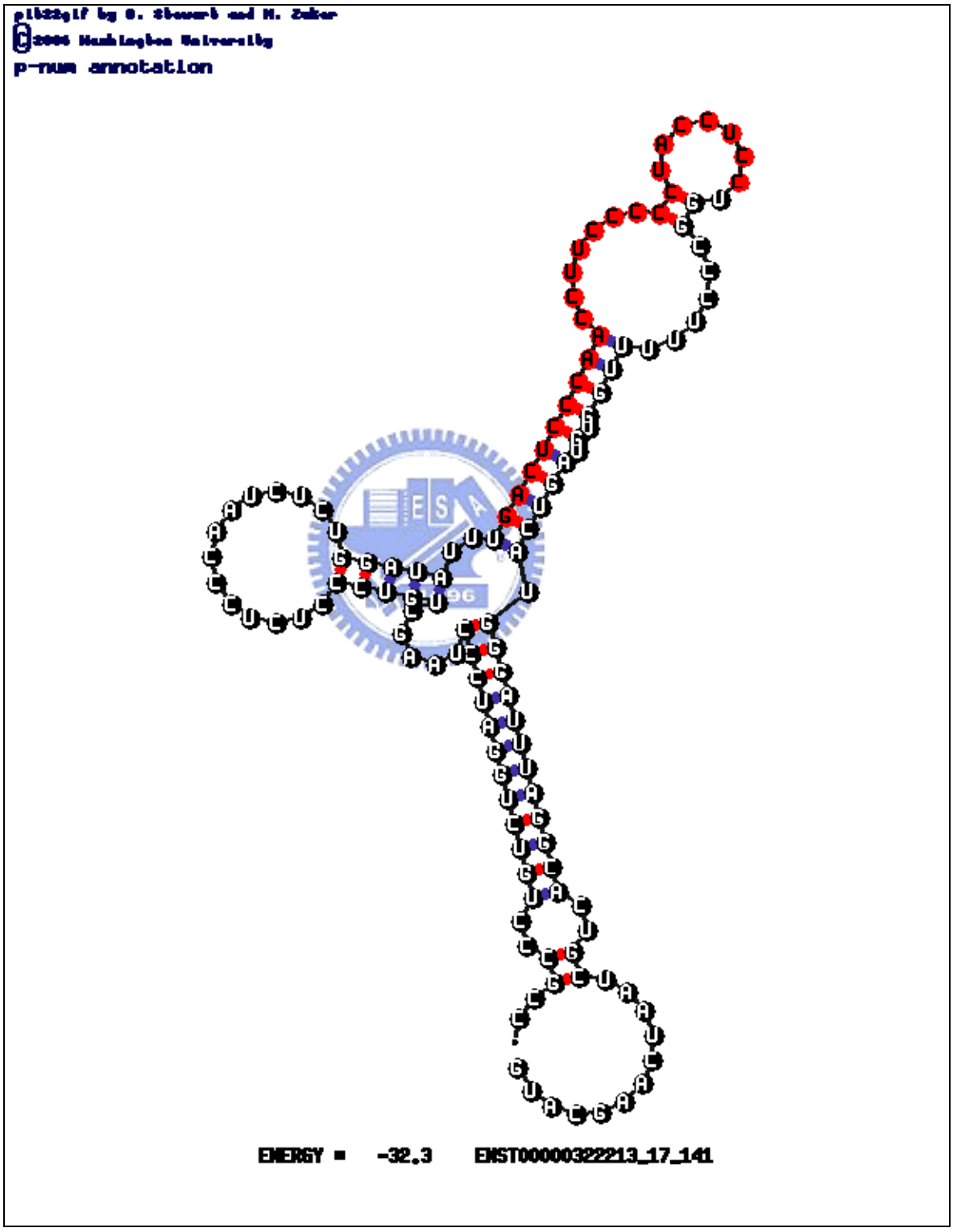


Figure 5.1 The original RNA secondary structure.

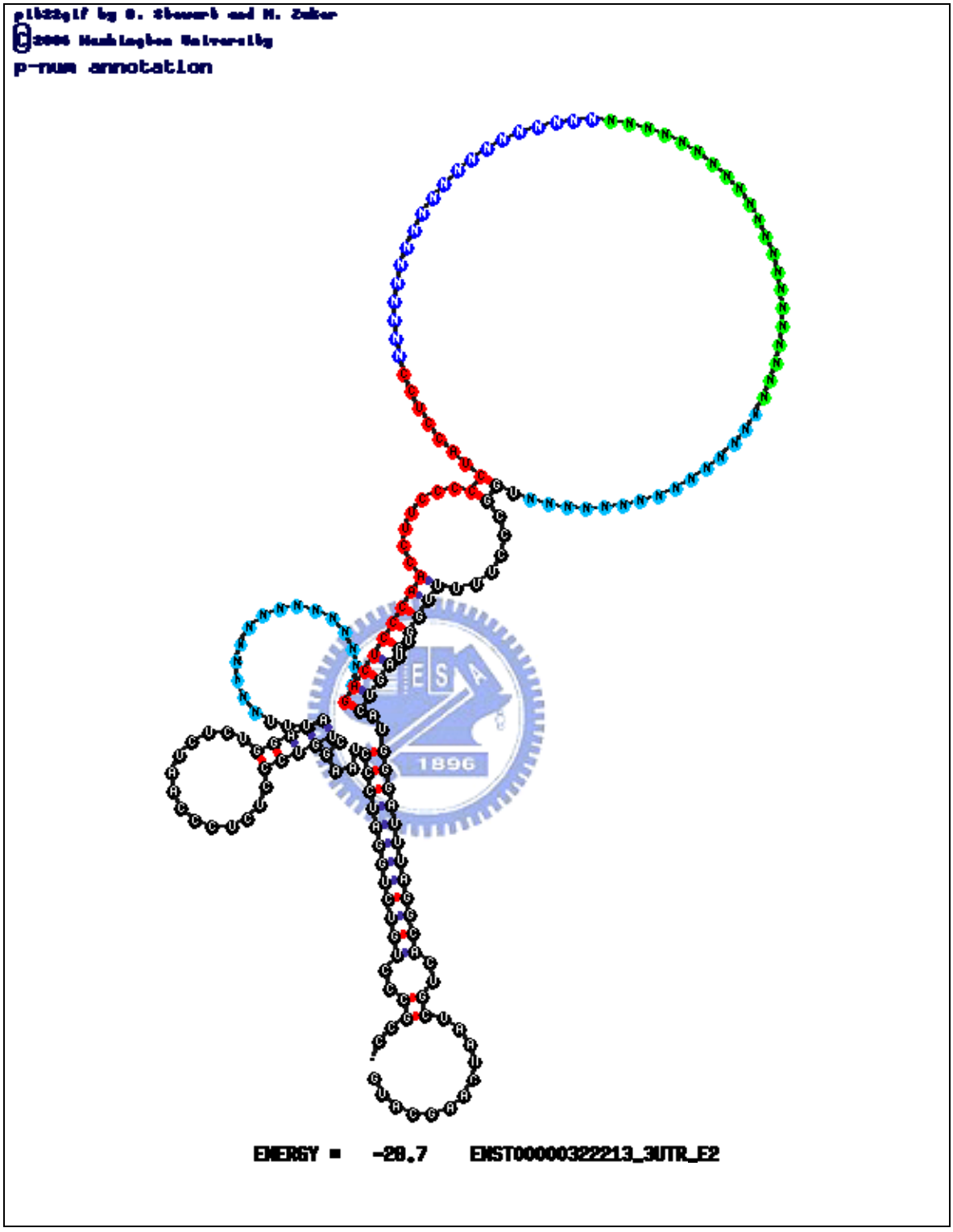


Figure 5.2 The RNA secondary structure after adding the linker.

5.3 Comparison to Other Tools

There are comparisons between miRTar and current released target prediction programs (Table 5.1).

	MicroInspector	miRU	miRanda	TargetScan	RNAhybrid	miRTar (Our method)
Species	arthropods, vertebrates, plants, nematodes, viruses	plants	Human Flies	Vertebrates	Flies	Vertebrates
Service type	Web server	Web server	Tool	Tool	Tool/Web server	Web server
Web interface with real-time response	Yes	Yes	No	No	Yes	Yes
Downloadable	No	No	Yes	Yes	Yes	No
Performance (seconds)	*	*	~108	~20	~804	~17
Provide miRNA/mRNA duplex structure	Yes	No	Yes	No	Yes	Yes

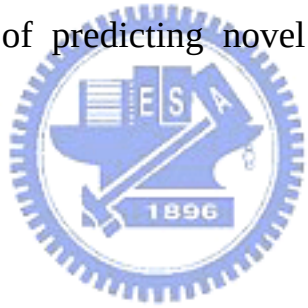
Table 5.1 Comparison between miRTar and current released target prediction programs.

5.4 Future Works

The improvements of our web system such as more species supporting, and response time reduction can be paid more attention. This work will also provide the gene expression profile of the miRNA target gene.

Chapter 6 Conclusion

This work proposes and implements a fast and efficient miRNA target prediction web server, namely miRTar, to identify miRNA targets against to the conserved sequence database. Several miRNA targets information such as the secondary structure of the mRNA containing the target site, GO annotation, and the conservation of target site linked to UCSC Genome Browser are provided in the web interface. This work uses several known cases of miRNA targets to evaluate our system, and the results suggest that our system is capable of predicting novel miRNA targets in an efficient manner.



References

1. Lau, N.C., et al., *An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 858-62.
2. Lee, R.C. and V. Ambros, *An extensive class of small RNAs in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 862-4.
3. Lagos-Quintana, M., et al., *Identification of novel genes coding for small expressed RNAs.* Science, 2001. **294**(5543): p. 853-8.
4. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation.* Nat Rev Genet, 2004. **5**(7): p. 522-31.
5. Elbashir, S.M., et al., *Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells.* Nature, 2001. **411**(6836): p. 494-8.
6. Elbashir, S.M., W. Lendeckel, and T. Tuschl, *RNA interference is mediated by 21- and 22-nucleotide RNAs.* Genes Dev, 2001. **15**(2): p. 188-200.
7. Ambros, V., et al., *MicroRNAs and other tiny endogenous RNAs in C. elegans.* Curr Biol, 2003. **13**(10): p. 807-18.
8. Kuwabara, T., et al., *A small modulatory dsRNA specifies the fate of adult neural stem cells.* Cell, 2004. **116**(6): p. 779-93.
9. Novina, C.D. and P.A. Sharp, *The RNAi revolution.* Nature, 2004. **430**(6996): p. 161-4.
10. Wightman, B., I. Ha, and G. Ruvkun, *Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans.* Cell, 1993. **75**(5): p. 855-62.
11. Wightman, B., I. Ha, and G. Ruvkun, *Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern-Formation in C-Elegans.* Cell, 1993. **75**(5): p. 855-862.
12. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome Biol, 2003. **5**(1): p. R1.
13. Allen, E., et al., *Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana.* Nat Genet, 2004. **36**(12): p. 1282-90.
14. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
15. Jover-Gil, S., H. Candela, and M.R. Ponce, *Plant microRNAs and development.* Int J Dev Biol, 2005. **49**(5-6): p. 733-44.
16. Wienholds, E. and R.H. Plasterk, *MicroRNA function in animal development.* FEBS Lett, 2005. **579**(26): p. 5911-22.
17. Yang, M., Y. Li, and R.W. Padgett, *MicroRNAs: Small regulators with a big impact.* Cytokine Growth Factor Rev, 2005. **16**(4-5): p. 387-93.

18. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II.* Embo J, 2004. **23**(20): p. 4051-60.
19. Lai, E.C., *microRNAs: runts of the genome assert themselves.* Curr Biol, 2003. **13**(23): p. R925-36.
20. Lai, E.C., C. Wiel, and G.M. Rubin, *Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes.* Rna, 2004. **10**(2): p. 171-5.
21. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.
22. Rehmsmeier, M., et al., *Fast and effective prediction of microRNA/target duplexes.* Rna, 2004. **10**(10): p. 1507-17.
23. Hsu, P.W., et al., *miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D135-9.
24. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
25. Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* Rna, 2005.
26. Zhang, Y., *miRU: an automated plant miRNA target prediction server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W701-4.
27. Rusinov, V., et al., *MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W696-700.
28. Griffiths-Jones, S., *The microRNA Registry.* Nucleic Acids Res, 2004. **32**(Database issue): p. D109-11.
29. Washietl, S., I.L. Hofacker, and P.F. Stadler, *Fast and reliable prediction of noncoding RNAs.* Proc Natl Acad Sci U S A, 2005. **102**(7): p. 2454-9.
30. Birney, E., et al., *Ensembl 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D556-61.
31. Hinrichs, A.S., et al., *The UCSC Genome Browser Database: update 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D590-8.
32. *The Gene Ontology (GO) project in 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D322-6.
33. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.
34. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Res, 2003. **31**(13): p. 3429-31.
35. Robins, H., Y. Li, and R.W. Padgett, *Incorporating structure to predict microRNA targets.* Proc Natl Acad Sci U S A, 2005. **102**(11): p. 4006-9.
36. Kloosterman, W.P., et al., *Substrate requirements for let-7 function in the developing zebrafish embryo.* Nucleic Acids Res, 2004. **32**(21): p. 6284-91.

37. Kiriakidou, M., et al., *A combined computational-experimental approach predicts human microRNA targets.* Genes Dev, 2004. **18**(10): p. 1165-78.
38. Hornstein, E., et al., *The microRNA miR-196 acts upstream of Hoxb8 and Shh in limb development.* Nature, 2005. **438**(7068): p. 671-4.