การระบุลักษณะเฉพาะของอาร์เอ็นเอเอโดยใช้ฐานข้อมูลการแสดงออกของยีน

นายนิรีกุมาร จายาวาดาร เจอมานี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

IDENTIFICATION OF RNAa CHARACTERISTICS USING GENE EXPRESSION OMNIBUS

Mr. Nilesh Gramani

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in

Computer Science and Information Technology

Department of Mathematics and Computer Science

Faculty of Science

Chulalongkorn University

Academic Year 2011

Thesis Title	Identification of RNAa Characteristics using Gene Expression Omnibus
By	Mr. Nilesh Gramani
Field of Study	Computer Science and Information Technology
Thesis Advisor	Assistant Professor Chatchawit Aporntewan, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

…….. Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

…….. Chairman
(Associate Professor Peraphon Sophatsathit, Ph.D.)

…….. Thesis Advisor
(Assistant Professor Chatchawit Aporntewan, Ph.D.)

…….. External Examiner
(Associate Professor Nachol Chaiyaratana, Ph.D.)

นิรีกุมาร จายาวาดาร เจอมานี : การระบุลักษณะเฉพาะของอาร์เอ็นเอเอโดยใช้ฐานข้อมูลการแสดงออกของยีน. (IDENTIFICATION OF RNAa CHARACTERISTICS USING GENE EXPRESSION OMNIBUS) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ.ดร.ชัชวิทย์ อาภรณ์เทวัญ, 42 หน้า.

การกระตุ้นด้วยอาร์เอ็นเอเป็นกระบวนการทางชีววิทยาที่ไมโครอาร์เอ็นเอจับกับโปรโมเตอร์ของยีนและทำให้ยีนทำงานมากขึ้น อาร์เอ็นเอเอเป็นกุญแจสำคัญสำหรับการรักษาการติดเชื้อไวรัสและโรคที่เกี่ยวกับเมแทบอลิซึม เช่น เอชไอวีและมะเร็ง อย่างไรก็ตามกระบวนการอาร์เอ็นเอเอส่วนใหญ่ยังไม่เป็นที่ทราบแน่ชัด และมีรายงานน้อยมาก การทำแล็บจะมีค่าใช้จ่ายและเสียเวลามาก แต่เราจะใช้การวิเคราะห์ด้วยคอมพิวเตอร์เพื่อจำแนกไมโครอาร์เอ็นเอที่อาจจะกระตุ้นการถอดรหัสโดยโปรตีนอาร์โกนอตทวู เราใช้ข้อมูลไมโครอาร์เรย์จาก Gene Expression Ombinus (GEO) เพื่อหายีนที่ทำงานเพิ่มขึ้นเมื่อเติมไมโครอาร์เอ็นเอ และทำงานน้อยลงเมื่อกำจัดอาร์โกนอตทวูออกไป ในขั้นสุดท้ายเราทำการปรับแนวระหว่างไมโครอาร์เอ็นเอและยีนโปรโมเตอร์โดยใช้ขั้นตอนวิธีสมิธวอเตอร์แมน เราพบไมโครอาร์เอ็นเออย่างน้อย 6 ตัวที่อาจจะจับบนโปรโมเตอร์และกระตุ้นการถอดรหัส

ภาควิชา ..คณิตศาสตร์และวิทยาการคอมพิวเตอร์.. ลายมือชื่อนิสิต .....

สาขาวิชา ..วิทยาการคอมพิวเตอร์.. ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก

..และเทคโนโลยีสารสนเทศ.. .....

ปีการศึกษา ..2554..

5273624223 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
KEYWORDS : microRNA, miRNA, Ago2, Argonaute2, RNA Activation, RNAa

NILESH GRAMANI : IDENTIFICATION OF RNAa CHARACTERISTICS USING GENE EXPRESSION OMNIBUS. ADVISOR : ASST. PROF. CHATCHAWIT APORNTEWAN, Ph.D., 42 pp.

RNA activation (RNAa) is a biological process characterized by miRNAs (microRNAs) binding to gene promoters and resulting in up regulation. RNAa is a key to cure virus infection and metabolic diseases such as HIV and cancer. However, the role of RNAa remains largely unknown. Very few cases have been reported in the past decade. In contrast to wet-lab approach which is costly and time consuming, we have performed a computer-based analysis to identify miRNAs that may activate transcription through Argonaute2 (Ago2). All microarray data are collected from Gene Expression Omnibus (GEO). We aim to find the genes that are commonly 1) up-regulated in a miRNA transfection experiment and 2) down-regulated in an Ago2 knockdown experiment. Finally we have performed local sequence alignment between the transfected miRNA and the gene promoters. Smith-Waterman (SW) algorithm is employed to find the best alignment score. Our findings show at least 6 miRNAs that could bind on promoters and may activate transcription.

Department : ..Mathematics and Computer Science.. Student's Signature……..……..……..
Field of Study: ..Computer Science and.. Advisor's Signature ……..……..……..
..Information Technology..

Academic Year : ..2011..

Acknowledgements

First, I would like to express my sincere acknowledgements to my advisor Asst. Prof. Chatchawit Aporntewan at the department of Mathematics in Chulalongkorn University, for support of my study and further research. His guidance helped me in all the time of research. I could not have imagined having a better advisor and instructor for my master study.

I would like to thank the rest of my thesis committee: Assoc. Prof. Peraphon Sophatsathit (Chair), Asst. Prof. Chatchawit Aporntewan (Advisor), Assoc. Prof. Nachol Chaiyaratana Committee (external), for their comments, encouragements, and hard questions. I am also grateful to Dr. Apiwat Mutirangura for enlightening me the first glance of research.

I thank my fellow classmate Guillermo Delgado from Spain and Chitphon Waithayanon from Thonburi (Thailand), for the project discussion in different subjects. We were working together before deadlines, and for all the fun we have had in the last 2 years.

Last but not the least; I would like to thank my parents, brother and my girlfriend: who encouraged me further study and supported me in financial problems.

I, Nilesh Gramani, dedicate this work to the entire people mentioned above. Without them, this work would never be done, I have faced many problems during this research, but as far as I concern, they were worthless against this incredible achievement.

Contents

Chapter Page

List of Tables

List of Figures

CHAPTER I

Introduction

Ribonucleic acid (RNA) is commonly found in organic cells. It is believed that RNA serves only as an intermediate in the transcription process. Now we know that RNA is a regulatory element that mediates gene expression. The discovery of RNA interference (RNAi), awarded Nobel Prize in Physiology in 2006, indicates that RNA plays an important role in suppressing the transcription [1]. RNAi is characterized by the binding of a small RNA to a messenger RNA (mRNA). The target mRNA is either degraded or destroyed. As a result, the corresponding gene is down-regulated. An important protein for RNAi activity is Ago2 which loads a short double-strand RNA (dsRNA) and rips it into two single-strand RNAs. The passenger strand is discarded. Ago2 takes only the guide strand, and then forms RNA-Induced Silencing Complex (RISC). This complex binds to a target mRNA whose sequence is complementary to the guide strand.

The opposite mechanism that activates transcription is referred to as RNA activation (RNAa) [2, 3]. RNAa is characterized by an increase of gene expression after introducing a small dsRNA which is complementary to the gene promoter. Ago2 may be required for RNAa activity because in an Ago2 knockdown experiment the gene expression level did not increase [4]. However, RNAa has been reported case by case, and the underlying process of RNAa remains largely unknown.

Although synthetic dsRNAs have been widely used in RNAi and RNAa studies, both of them are natural mechanisms. In nature, dsRNAs called microRNA (miRNA) are ubiquitous in eukaryotic cells. miRNAs are encoded in DNA like other genes, but they are non-coding and are not translated to functional proteins. Thus, the role of miRNAs is solely a regulatory element. The regulatory functions of miRNA through RNAi have been studied extensively, but very few miRNAs that involve RNAa have been reported.

This thesis aims to search for miRNAs that may activate transcription and increase gene expression. Our strategy employed a public gene expression database for screening the activated genes. Then, we performed local sequence alignment between miRNAs and the promoters of those genes.

1.1 Objectives

- To investigate RNA activation (RNAa) in a public gene expression database, called Gene Expression Omnibus (GEO) [5].
- To identify miRNAs that may up-regulate genes through RNAa in cancers.
- To identify miRNA binding sites on gene promoters using sequence alignment.
- To select candidate miRNAs and genes for further verification in web lab.
- To identify characteristics of RNAa such as sense/antisense and distance from TSS.

1.2 Scope of the Work

- We have selected miRNA transfection experiments and Ago2 KO experiments from a public database called Gene Expression Omnibus (GEO). Note that transfection is the process of deliberately introducing nucleic acids into cells.
- We intersect two microarray experiments using our software called “CU-DREAM” [6].
- The local sequence alignment used in this thesis is based on Smith-Waterman algorithm [7].
- We focus only on Homo sapiens.

1.3 Expected Outcome

- A number of miRNAs that may up-regulate genes through RNAa.
- A number of candidate genes that are targets of RNAa.
- The characteristics of RNAa, e.g. binding length, distance from transcription start site (TSS), etc.

1.4 Problem Formulation

Our research problem is formulated in the following steps.

- Firstly, we will intersect a miRNA transfection experiment with an AGO2 knockdown experiment. Through the intersection, a number of genes will be counted in both experiments. The p-value and odds ratio will be measured that these two datasets significantly share Up-Down regulated genes, so we can determine that those genes in set *A* are candidate genes for RNAa.
- Secondly, we will collect miRNA sequence from mirBase [8]. To find miRNA binding site, we will align transfected miRNA to gene promoters through local sequence alignment, and then we will identify those miRNAs that may activate transcription through Ago2. We will employ Smith-Waterman algorithm for local sequence alignment. SW Algorithm aims to find the best alignment score. Through the local sequence alignment we can putatively identify binding location of those transfected miRNAs.
- Thirdly, we will find the characteristics of RNAa that hypothetically mediate the process of RNAs. For example,
    - Location of the binding sites: distance from the transcriptional start sites (TSS).
    - Orientation: sense or antisense
    - Binding: perfect or not perfect, binding length
    - miRNA and DNA (Gene Promoter): common sequence, etc.

CHAPTER II

Theoretical Background

This chapter will explain basic knowledge of human genome and some cellular processes that are important to understand the rest of thesis.

2.1 DNA & RNA

There are two types of nucleic acids that are found in all living cells: DNA and RNA. Deoxyribonucleic Acid (DNA) is most probably found in nucleus of cell while Ribonucleic Acid (RNA) is found in cytoplasm. DNA contains genetic information that is stored as codes to make RNA molecule and then RNA molecule converts into proteins.

DNA is basically a long molecule chain that contains genetic code of life. The code is made up with four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA determines the features of living things, for example, eye and hair color. The four bases are paired with a strict rule. Base A can bind only with base T, and base C can bind only with base G. Each base pair is connected with sugar and phosphate molecules called a nucleotide. Nucleotides are arranged in two long strands with different base pairs to form the structure of DNA called double helix.

Ribonucleic acid (RNA) is basically found in the cytoplasm of the cell. RNA is also nucleic acids that contains ribose (sugar), phosphate and four chemical bases; adenine (A) uracil (U) cytosine (C) and guanine (G). RNA is a molecule similar to DNA but different structure. RNA is a single strand and thymine (T) is replaced with uracil (U). All components ribose, phosphate and four bases are associated with the control of cellular chemical activities. RNA transmits genetic information from DNA to make functional proteins.

Different types of RNA exist in the living cell: messenger RNA (mRNA) ribosomal RNA (rRNA) and transfer RNA (tRNA). More recently, some small interfering RNAs (siRNAs) have been found to be involved in regulating gene expression or short RNAs (shRNAs). Both RNA and DNA structures are shown in Figure 1.

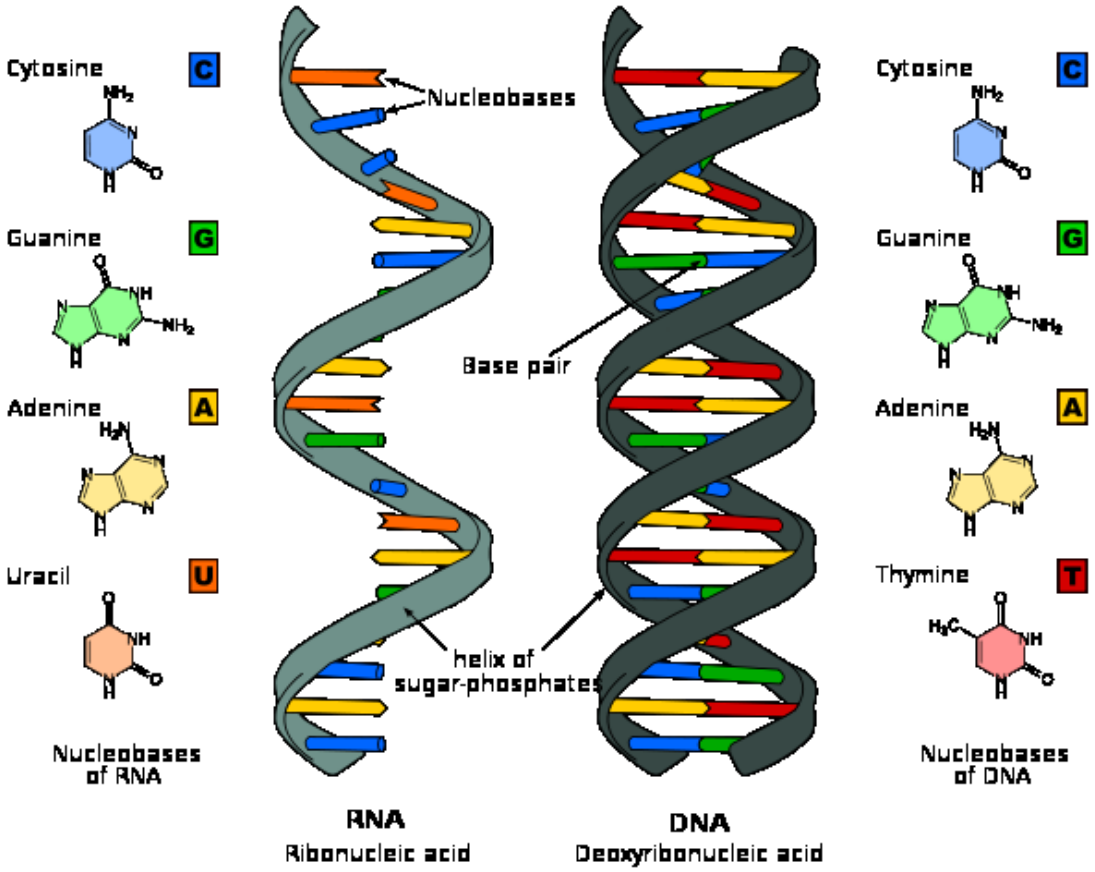

Figure 1: Ribonucleic Acid (RNA) and Deoxyribonucleic Acid (DNA).

2.2 Genes

Two DNA strands form a chromosome (see Figure 2). In a human cell, there are a total of 23 chromosomes. A gene is a coding region on DNA. The coding region or gene body is the part of DNA that can be transcribed into mRNA and then translated into a protein. The gene body is next to its promoter, which is the sequence that binds to transcriptional factors and allows transcription process. In humans, there are about 30,000 genes. A gene produces a corresponding protein. Subsequently, a protein (or a gene) determines a trait, for instance, hair color. However, most traits are complex and involve multiple genes.

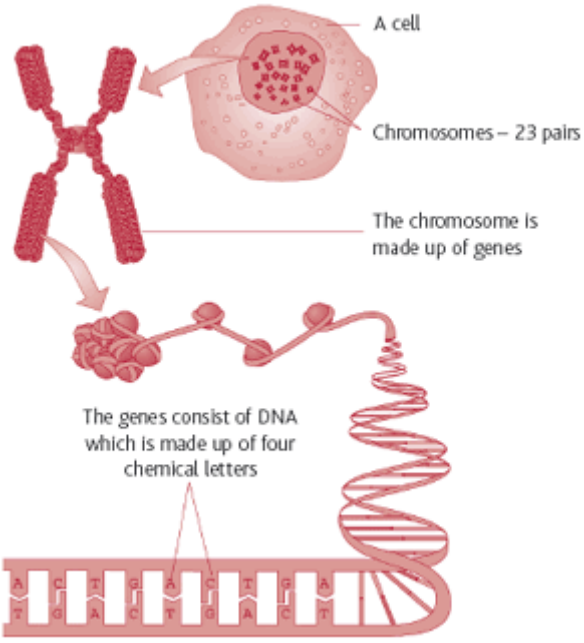

Figure 2: Chromosomes and genes.

2.3 Transcription Process

Transcription is a biological process that makes RNA molecule from DNA (or a gene). The transcription process is depicted in Figure 3. Firstly, RNA polymerase identifies a specific base sequence on a gene promoter and binds to it. Secondly, RNA polymerase unwinds the DNA and starts producing RNA that is complementary to the template strand. Note that RNA is made of uracil (U) instead of thymine (T). Finally, the termination code at the end of gene tells RNA polymerase to stop transcription. Next, the translation process converts RNA to proteins (see Figure 4).

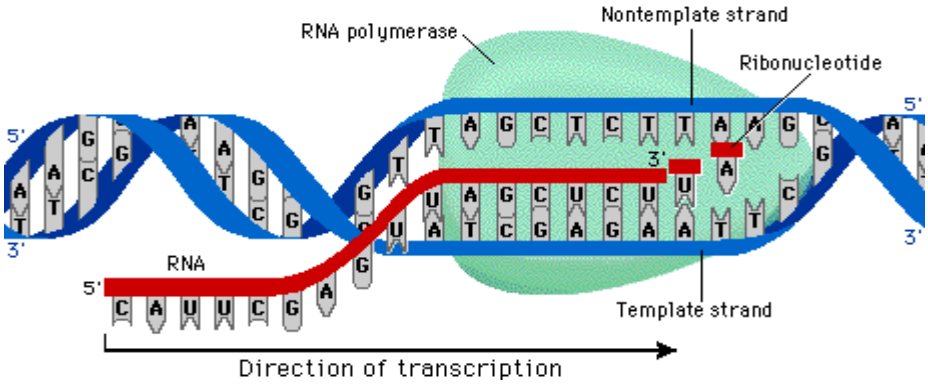

Figure 3: Transcription process.

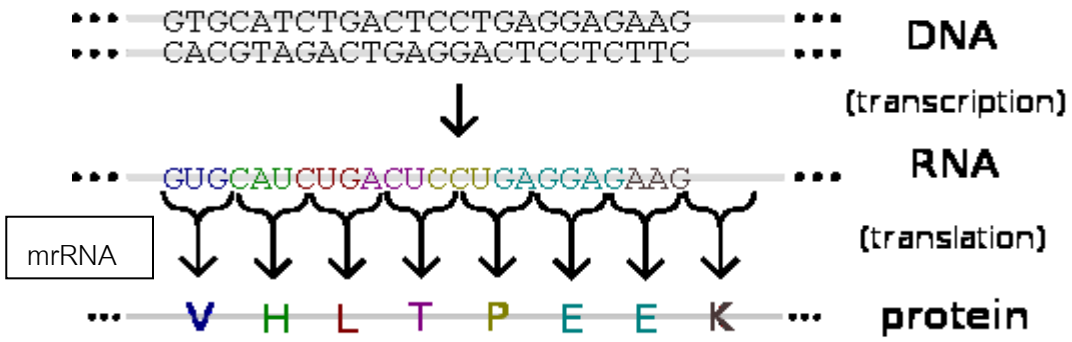

Figure 4: Gene expression.

Translation is a biological process that messenger RNA (mRNA) is converted into functional proteins. Moreover, Translation is the third stage of overall process of gene expression. After the transcription process a ribosome is an essential component of cells that collects the twenty specific amino acid molecules to form the particular protein molecule and it is also determined by the nucleotide sequence of an RNA molecule. Generally DNA sequence in genes is copied into messenger RNA. (because in general case DNA makes RNA and RNA makes proteins). Ribosome can read the information in mRNA and uses them to convert to a particular protein. Ribosome translates the genetic information from the mRNA into proteins. Ribosomes do this by binding to an mRNA and using it as a template for determining the correct sequence of amino acids in a particular protein. For instance, A binds to U and C binds to G. The amino acids are attached to transfer RNA (tRNA) molecules, which enter one part of the ribosome and bind to the messenger RNA sequence. The ribosome starts matching tRNA antisense sequences to the mRNA sense sequence. When the ribosome reaches one of the "stop" codes, for example when the A side of the ribosome faces a stop codes (For instance, UAA, UAG, or UGA are faced error in coding between tRNA and mRNA). So tRNA cannot identify or bind continuously, and then the ribosome releases both the newly born protein chain and the mRNA. These newly born protein chain acts as a functional protein in a cell. The translation process is clearly illustrated in Figure 5.

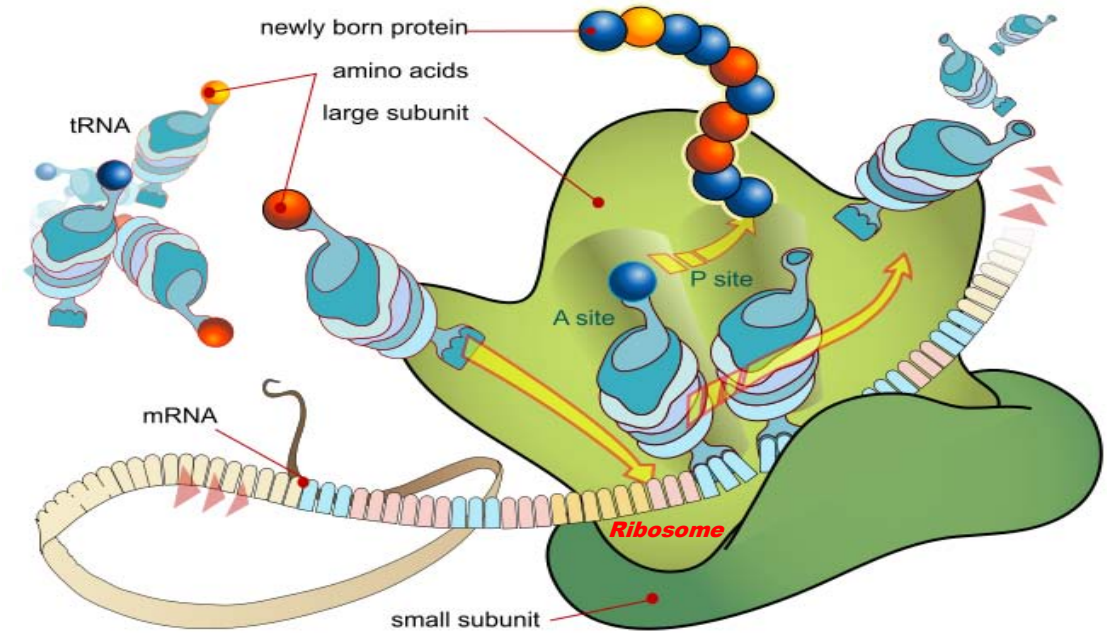

Figure 5: Translation process.

2.4 Microarray and Gene Expression

Gene expression is referred to the amount of mRNA produced in the transcription process. To measure the gene expression, we use microarray. A microarray chip contains a thousand of probes. Each probe binds to a specific mRNA. The main idea is to color mRNA and take a photo. The intensity of the color is proportional to the amount of mRNA produced. A common technique to compare gene expression between an experimental group and a control group is to color them with different colors, red and green. Next, both group are poured together and taken a photo. The same amount of gene expression in both group yields yellow, which is a combination of red and green. A photo of two-color experiment is shown in Figure 5.

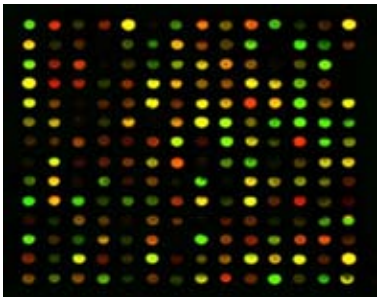

Figure 6: A microarray chip.

2.5 RNA Interference

RNA interference (RNAi) is characterized by the binding of a small RNA to a messenger RNA (mRNA). The target mRNA is either degraded or destroyed. As a result, the corresponding gene is down-regulated (see Figure 6). RNAi process begins with a short double strand RNA (dsRNA), about 22 – 25 nucleotide long. Next, a protein called Dicer chops dsRNA into smaller pieces, and then loaded into RNA-induced silencing complex (RISC). Only the guide strand remains, the passenger strand is discarded. Note that the formation of RISC requires Ago2 protein. Subsequently, the RISC binds on mRNA of which the sequence is complementary to the guide strand. A perfect match results in mRNA cleavage (completely disrupted), but an imperfect match only degrades mRNA. Hence, RNAi lowers expression level of target genes.

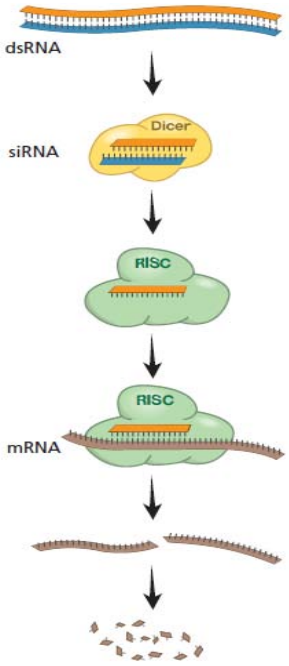

Figure 7: The process of RNAi.

2.6 RNA Activation

RNA Activation (RNAa) is the opposite process of RNAi. RNAa is characterized by an increase of gene expression after inserting a small dsRNA [2, 3, 4] (see Figure 7). It is believed that the inserted dsRNA binds on gene promoter and

amplifies the transcription. It is also believed that Ago2 is required for RNAa because RNAa is not observed in Ago2 knockout experiments. However, very few cases and very few spots on the whole genome have been reported. The general process of RNAa remains largely unknown.

The terms “small interfering RNA (siRNA)” and “small activating RNA (saRNA)” are coined to name small RNA by its effect. In fact, there are two sources of small RNA: synthetic and natural RNA. Small RNA can be synthetically made to perfectly match on a target. On the other hand, microRNA (miRNA) is natural and found ubiquitously in living cells. In the past, it was believed that miRNA has no functions. Now it is accepted that miRNA acts as a regulatory element that mediates gene expression [9] (see Figure 7).

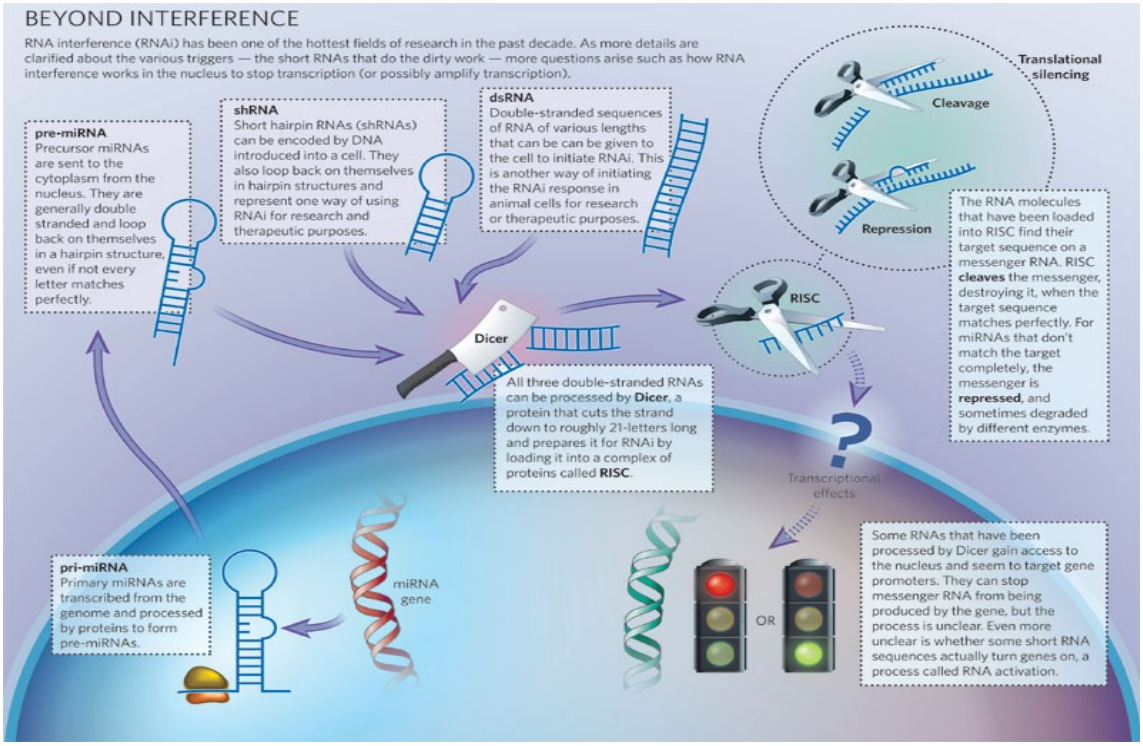

Figure 8: The process of RNAa.

RNAa

2.7 Local Sequence Alignment

Sequence alignment is a method for finding the similarity between two sequences. The sequences may be DNA, RNA, or proteins. The sequence alignment

allows some stretching or shrinking on the sequences to obtain the best alignment score. Our aim is to perform local alignment between a small RNA and a gene promoter (RNA vs. DNA). A well-known local sequence alignment is Smith-Waterman algorithm [7]. SW algorithm defines the best alignment score as follows.

$$E_{i,j} = max\begin{Bmatrix} E_{i,j-1} - G_{ext} \\ H_{i,j-1} - G_{init} \end{Bmatrix}$$

$$F_{i,j} = max\begin{Bmatrix} F_{i-1,j} - G_{ext} \\ H_{i-1,j} - G_{init} \end{Bmatrix}$$

The alignment score $H_{i,j}$ where $1 \leq i \leq m$ and $1 \leq j \leq n$ is defined by the following equation.

$$H_{i,j} = max\begin{Bmatrix} 0 \\ E_{i,j} \\ F_{i,j} \\ H_{i-1,j-1} + W(q_i, d_j) \end{Bmatrix}$$

The value for $E_{i,j}$ $F_{i,j}$ and $H_{i,j}$ are equal to 0 when $i = 0$ or $j = 0$. $Q = q_1, \dots, q_m$ and $D = d_1, \dots, d_n$ are the two sequences to be aligned. The lengths of Q and D are equal to m and n respectively. $H_{i,j}$ is the best alignment score using only the first i characters of Q and the first j characters of D. $E_{i,j}$ and $F_{i,j}$ are opening a gap on D and Q respectively. $W(q_i, d_j)$ is the score of matching characters $q_i$and d_j, defined by the scoring matrix in Table 1.

Table 1: A scoring matrix for DNA-RNA binding.

		RNA			
		A	U	C	G
DNA	A	0	2	0	0
	T	2	0	0	1
	C	0	0	0	2
	G	0	1	2	0

In our case, Q is DNA sequence (composed of A, T, C, G) and D is RNA sequence (composed of A, U, G, T). Scoring “2” indicates strong chemical bonds, scoring “1” indicates weak bonds, and “0” indicates no bonds. G_{init} is the penalty for

opening a gap, and G_{ext} is the penalty of extending gap by one character. Here is an example of aligning.

```
DNA = AGTGGCTCATGGCTCACACCTGAAATCCTAGCACTTTGGGAGGCCAAGGCAGG
RNA = UGUGGGGUUUUAGCUUCGUGAAG
```

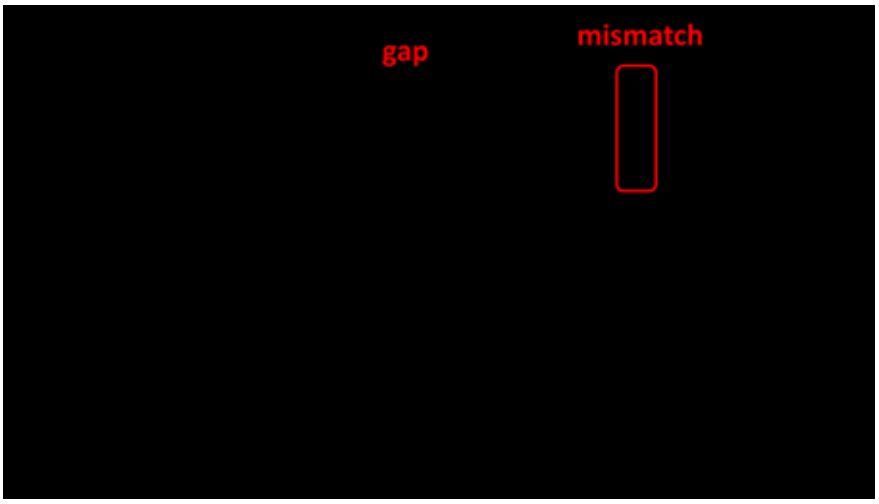

Figure 9: An alignment between DNA and RNA.

The main aim of sequence alignment is to match subsequences as far as possible. In above Figure 8, sequence 1 (DNA sequence) composed of A, T, C, and G and sequence 2 (RNA sequence) composed of (A, U, C, and G). The score of matching letters is defined by the DNA-RNA scoring matrix (see Table 1), for instance, A-U, T-A, C-G and G-C has strong chemical bound so we have given 2 points. T-G and G-U were weak bound so we have given only 1 point. The penalties of opening a gap and extending a gap are defined set at the same value, -2, while scoring 0 indicates that there is no bound or mismatch. In above example there are seventeen matches indicating with vertical bars and open gap and mismatch indicating by blank or space. According to scoring matrix, there are two mismatches can be identified; “C“ in sequence 1 with aligned “C” in sequence 2. Same as “T“ in sequence 1 with aligned “U” in sequence 2. Space is indicating gaps in the sequence 2. Open gap is important to get good alignment for rest of the letters between both sequences to get a good alignment score.

The best alignment score is obtained by filling the dynamic programming table in Figure 10 from top to bottom and from left to right. The maximum number in the table is the best alignment score (not necessarily to be the number at the lower-right corner). Once the alignment is done, a score is counted to each letters called base pair (bp). For the above example, the best alignment score is [(17) Strong match x (2) Score] + [(3) weak match x (1) Score] + [(2) mismatch x (0) Score] + [(1) open gap x (-2) Penalties + [(0) extend gap x (0) Penalties = [34+3+0+ (-2) +0] = 35. The markup line is obtained by tracing back the lower-right corner to the upper-left corner. An arrow indicates the choice of matching, opening a gap, or extending a gap.

-	DNA	C	A	C	A	C	C	T	G	A	A	A	T	C	C	T	A	G	C	A	C	T	T	T
RNA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	2	0	2	0	2	2	1	0	0	0	0	1	2	2	1	0	0	2	0	2	1	1	1
U	0	0	4	2	4	2	2	2	2	2	2	2	0	1	2	2	3	1	0	4	2	2	1	1
G	0	2	2	6	4	6	4	3	2	2	2	2	3	2	3	3	2	3	3	2	6	4	2	2
G	0	2	2	4	6	6	8	6	4	2	2	2	3	5	4	4	3	2	5	3	4	7	5	3
G	0	2	2	4	4	8	8	9	7	5	3	2	3	5	7	5	4	3	4	5	5	5	8	6
G	0	2	2	4	4	6	10	9	9	7	5	3	2	5	7	8	6	4	5	4	7	6	6	9
U	0	0	4	2	6	4	8	10	10	11	9	7	5	3	5	6	10	8	6	7	5	7	6	7
U	0	0	2	4	4	6	6	8	11	12	13	11	9	7	5	5	8	11	9	8	7	5	7	6
U	0	0	2	2	6	4	6	6	9	13	14	15	13	11	9	7	7	9	9	11	9	7	5	5
U	0	0	2	2	4	6	4	4	7	11	15	16	15	13	11	9	9	8	9	11	11	9	7	5
A	0	0	0	2	2	4	4	6	5	9	13	15	18	16	14	13	11	9	8	9	9	13	11	9
G	0	2	0	2	2	4	6	5	6	7	11	13	16	20	18	16	14	12	11	9	11	11	14	12
C	0	0	2	0	2	2	4	6	7	6	9	11	14	18	20	18	16	16	14	12	10	11	12	14
U	0	0	2	2	2	2	2	4	7	9	8	11	12	16	18	20	20	18	16	16	14	12	11	12
U	0	0	2	2	4	2	2	2	5	9	11	10	11	14	16	18	22	21	18	18	16	14	12	11
C	0	0	0	2	2	4	2	2	4	7	9	9	10	12	14	16	20	24	22	20	18	16	14	12
G	0	2	0	2	2	4	6	4	2	5	7	9	10	12	14	15	18	22	26	24	22	20	18	16
U	0	0	4	2	4	2	4	6	5	4	7	9	9	10	12	14	17	20	24	28	26	24	22	20
G	0	2	2	6	4	6	4	5	6	5	5	7	10	11	12	13	15	18	22	26	30	28	26	24
A	0	0	2	4	6	4	6	6	5	6	5	5	9	10	11	14	13	15	20	24	28	32	30	28
A	0	0	0	2	4	6	4	8	6	5	6	5	7	9	10	13	14	13	18	22	26	30	34	32
G	0	2	0	2	2	6	8	6	8	6	5	6	6	9	11	11	13	14	16	20	24	28	32	35

Figure 10: A dynamic programming table for sequence alignment.

2.8 Basic Statistics

2.8.1 Odds Ratio

Odds Ratio (OR) is a statistical test for comparing whether the probability of certain cases or events is the same for two groups or not. The OR is defined below in Table 2.

Table 2: Odds Ratio.

	Case group	Control group	Odds
Event happens	a	b	a / b
Event does not happen	c	d	c / d
Total	a + c	b + d	**OR = (a/b) / (c/d)**

As shown in the table above, the odd of first row (event happens) is a/b, and the odd of the second row (event does not happen) is c/d. The odds ratio (OR) is simply the ratio of the two odds, which is (a/b) / (c/d) or (a x d) / (b x c). The interpretation of OR is as follows.

OR = 1, the event happens equally in both groups.

OR > 1, the event is more likely to happen in the case group.

OR < 1, the event is more likely to happen in the control group.

Suppose If we consider the data take from cancer study to determine whether the use of new treatment is better option for patients who are suffering from cancer, rather than normal treatment or not? The data consist of a sample of n individuals – are arranged in the 2 x 2 contingency table in Table 3.

Table 3: A 2x2 contingency table in cancer study.

	Died	Survived	Odds
Normal Treatment	a = 150	b = 250	a / b = 0.6
New Treatment	c = 20	d = 100	c / d = 0.2
Total	a + c = 170	b + d = 350	**OR = 0.6 / 0.2 = 3**

From above example, First we count the odds ratio (OR) that,

$$OR = \frac{\text{Died in Normal Treatment (a) / Survived in Normal Treatment (b)}}{\text{Died in New Treatment (c) / Survived in New Treatment (d)}}$$

$$OR = \frac{150/250}{20/100} = 3$$

Secondly, we will have to compute a confidence interval (CI of 95%) for odds ratio whether we are 95% confident that the new treatment covers the true proportion of individuals' samples rather than the normal treatment. Typically, we can count with log scale and be insured that sample of size (n) is sufficiently large. In addition, it is expected that the value of each selected entry in above contingency table should be at least 5. We are very well know that 95% of below example possible outcome lie between lower (-1.96) and upper (+1.96). The expression for a 95% CI for the natural logarithm of the OR is,

$$e^{In(OR)\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

Before we compute a CI for OR, we will have to measure normal standard error of size in contingency table in each category.

Standard error = $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{150}+\frac{1}{250}+\frac{1}{20}+\frac{1}{100}}$

For the data examining the relationship between people died and survived in cancer against both treatment, so the log of the estimated odd ratio is,

$$In\ (OR) = In(3) = 1.10$$

To find a 95% CI for the odds ratio itself, we will have to take antilogarithm of the both upper and lower limits from the below equation,

$$e^{In(OR)-1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}, e^{In(OR)+1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

So, a 95% confidence interval for the log of the odds ratio in both outlines lower and upper is,

$$e^{In(3)-1.96\sqrt{\frac{1}{150}+\frac{1}{250}+\frac{1}{20}+\frac{1}{100}}}, e^{In(3)+1.96\sqrt{\frac{1}{150}+\frac{1}{250}+\frac{1}{20}+\frac{1}{100}}}$$

$$e^{0.5708}, e^{1.6292}$$

$$(1.77, 5.11)$$

So further results we are 95% confident to determine that odds ratio more than 1 would be mean that normal treatment is more likely to kill patients. In other words, new treatment tends to save more patients between 1.77 and 5.11 times compared to the normal treatment. And more importantly, the new treatment reduces the death rate by 3 in comparison to the normal treatment.

2.8.2 Chi-Square Test

Table 4: Observed frequency.

	Died	Survived	Total
Normal Treatment	a = 150	b = 250	a + b = 400
New Treatment	c = 20	d = 100	c + d = 120
Total	a + c = 170	b + d = 350	n = 520

To examine the effectiveness of new treatment in above example, we want to know whether there is an association between the survival and treatments. Generally in Chi-square test investigators predict the scientific experiments start with the null hypothesis, and compare a p–value with α (null hypothesis), If p-value < α they reject the null hypothesis , for instance in the above example null hypothesis is,

H_0 There is no any significant difference between normal treatment and new treatment.

Against the alternative scenario is,

H_a There is significant difference between normal treatment and new treatment.

First, we have to calculate the expected frequency for each cell of the contingency table. Under the null hypothesis (H_0) we can determine that it should be rejected because if we count independent experiments in normal treatment and new treatment, the results will be identical like 50-50 in both treatments. So we will have to reject two independent categories and count the all 520 singular homogeneous sample. For Instance,

Overall amount of normal treatment is,

$\frac{400}{520} = 77\%$ Where $400 = (\ a\ +\ b\ ) = 150 + 250$ and

$520 = (\ a\ +\ b\ + c + d) = n$.

Overall amount of new treatment is,

$\frac{120}{520} = 23\%$ Where $120 = (\ c\ +\ d\ ) = 20 + 100$ and

$520 = (\ a\ +\ b\ + c + d) = n$.

As above results, 170 people died, so we expect that 77% of them are in normal treatment.

$$(a + b\ ) \times \frac{(a + c\ )}{n}\ = (150 + 250\ ) \times \frac{(150 + 20\ )}{520} = 130.77$$

And 23% of them are in new treatment.

$$(c + d\ ) \times \frac{(a + c\ )}{n}\ = (20 + 100\ ) \times \frac{(150 + 20\ )}{520} = 39.23$$

Second, 350 people survived, so we expect that 77% of them are in normal treatment.

$$(a + b\ ) \times \frac{(b + d\ )}{n}\ = (150 + 250\ ) \times \frac{(250 + 100\ )}{520} = 269.23$$

And 23% of them are in new treatment.

$$(c+d)\times\frac{(b+d)}{n}=(20+100)\times\frac{(250+100)}{520}=80.77$$

The above calculation is shown in Table 5.

Table 5: Expected frequency.

	Died	Survived	Total
Normal Treatment	(a + b) (a + c) / n = 130.77	(a + b) (b + d) / n = 269.23	a + b = 400
New Treatment	(c + d) (a + c) / n = 39.23	(c + d) (b + d) / n = 80.77	c + d = 120
Total	a + c = 170	b + d = 350	n = 520

$$\mathrm{X}^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}$$

Where,

X^2 = Chi-square test statistics

O_i = an observed frequency in each category

E_i = an expected (theoretical) frequency, asserted by the null hypothesis

n = the number of cells in the table.

Put all variables in the above chi-square equation the result would be shown as below.

$$\mathrm{X}^2=\frac{(150-130.77)^2}{130.77}+\frac{(20-39.23)^2}{39.23}+\frac{(250-269.23)^2}{269.23}+\frac{(100-80.77)^2}{80.77}$$

$$\mathrm{X}^2 = 2.68 + 8.94 + 1.30 + 4.34 = \mathbf{18.21}$$

For $\mathrm{X}^2 = 18.21$ distribution with 1 degree of freedom, the p-value is less than 0.001 (see Table 6), indicating statistical significance. The result is P-value < α (null hypothesis) concluded that observed case from the null hypothesis is more

significant. So we can determine that relative amount of individual cases in normal treatment and new treatment, the people who involved in cancer and who are curing a new treatment is more effective than normal treatment.

Table 6: Chi-square threshold of significance.

Degrees of freedom (df)	χ^2 value [11]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Nonsignificant								Significant		

It is important to note that Chi-square test is an approximation. However, it is a good approximation as long as the numbers in the 2x2 table are large. If the numbers are small, Fisher's exact test is more suitable.

CHAPTER III

Materials and Method

All data we have used in this thesis is from a public database called Gene Expression Omnibus (GEO). We use our software called CU-DREAM for intersecting two microarray experiments. The local sequence alignment is implemented by writing a small piece of code.

3.1 Gene Expression Omnibus (GEO)

The National Center for Biotechnology Information (NCBI) provides biomedical and genomic information through NCBI website. Gene Expression Omnibus (GEO) is a public database of gene expression experiments. At present, there are approximately half a billion samples of over 100 living organisms. These huge volumes are on-line resources that we can effectively browse, visualize and download from http://www.ncbi.nlm.nih.gov/geo/.

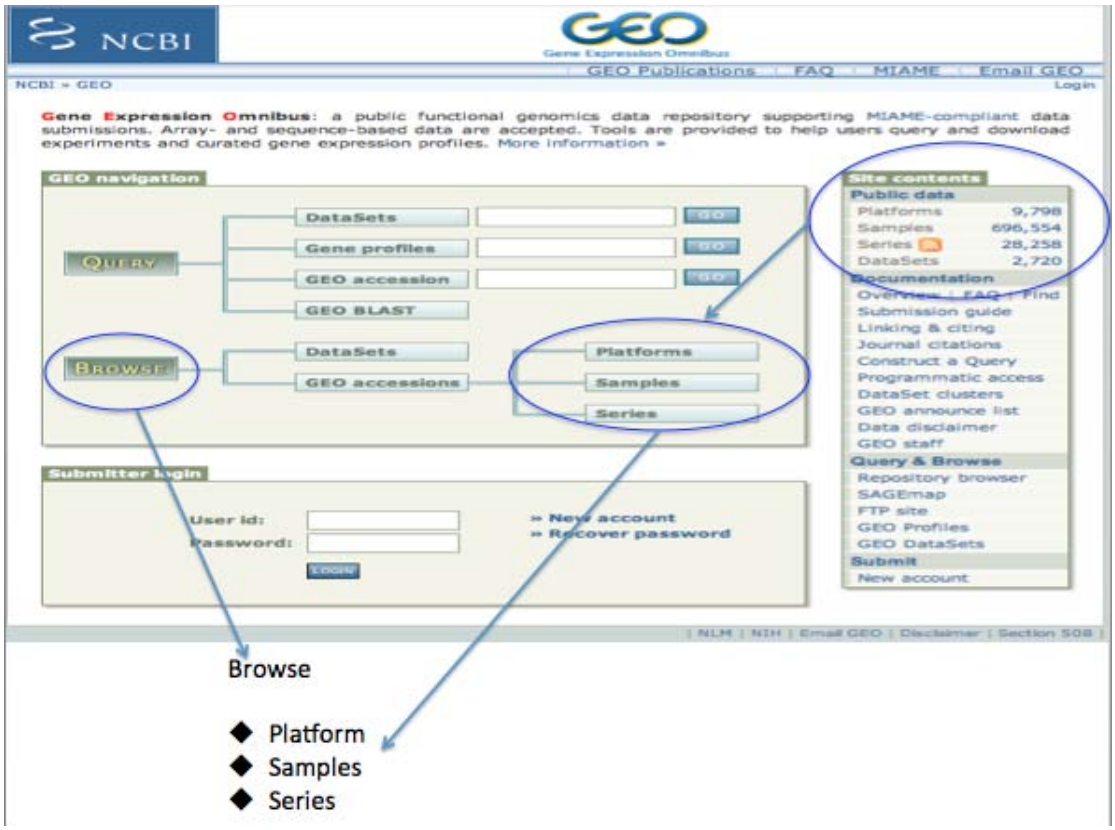

Figure 11: NCBI and GEO website.

GEO consists of 4 kinds of data as depicted in Figure 10.

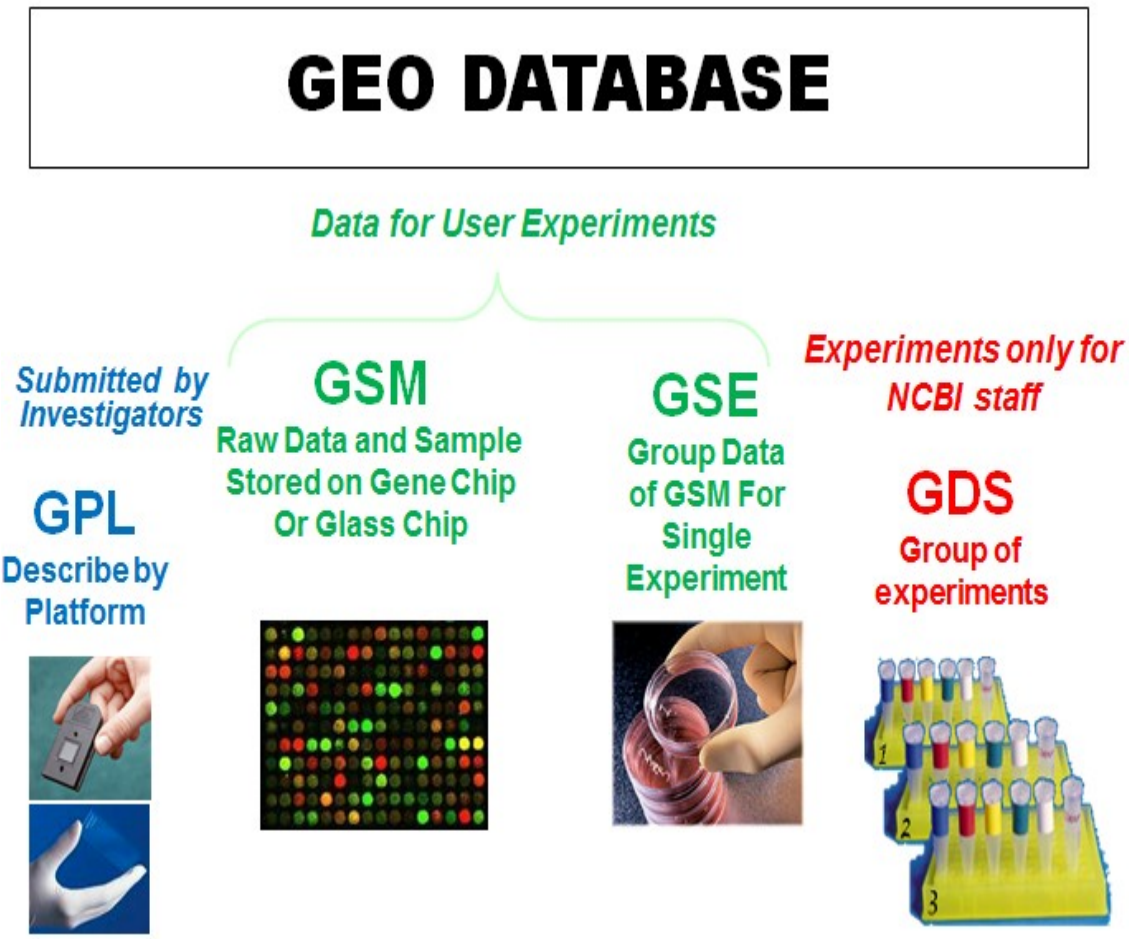

Figure 12: Four kinds of data on GEO.

1. Platform (GPL)

The Platform describes the physical setup of the analysis. It is analyzing specific product, for example, gene chip of cDNA (complementary DNA). GEO Platform is described by GPL. Platform is submitted by manufacturers.

2. Sample (GSM)

Samples are the individual array measurements. It describes biological materials and experimental conditions under which the sample was handled. In GEO, sample is described by GSM. Sample is submitted by experimentalists (users).

3. Series (GSE)

Series are sets of related samples, which considered a group of slide (chip data) to get single experiment. Series is described by GSE on GEO and also submitted by users.

4. Raw data (GDS)

These data set consist with original microarray scan images and describe by GDS on GEO, but GDS data are assembled by only NCBI (National Center for Biotechnology Information) GEO Staff.

We have selected microarray experiments that injected in cell single or multiple miRNAs and Ago2 knockdown experiments from GEO. We have used them for studying RNA activation through miRNAs.

Table 7: Micro RNA transfection and Ago2 experiments selected from GEO.

GSE (Series)	Transfected miRNAs	Title
GSE6207	miR-124	miR-124 transfection time course
GSE7864	miR-34a,34b,34c	A microRNA component of the p53 tumor suppressor network
GSE11701	miR-205	Genes modulated by miR-205 in DU145 prostate cancer cells
GSE13105	miR-192,215	Coordinated regulation of cell cycle transcripts by p53-inducible microRNAs, miR-192 and -215
GSE13296	miR-155	miR-155 KO in human dendritic cells
GSE16568	miR-22	Gene expression analyses of mir-22 overexpression in ovarian clear cell cancer cell line
GSE16569	miR-30a,30d	Gene expression analyses of mir-30a knockdown in ovarian clear cell cancer cell line
GSE16571	miR-100	Gene expression analyses of mir-100 overexpression in ovarian clear cell cancer cell line
GSE16572	miR-182	Gene expression analyses of mir-182 knockdown in ovarian clear cell cancer cell line
GSE16700	miR-31	Gene expression analyses of mir-31 overexpression in ovarian serous cancer cell line

GSE16908	miR-31	miR-31 inhibits lymphatic lineage-specific differentiation in vitro and lymphatic vessel development in vivo
GSE18510	miR-193b	miR-193b represses cell proliferation and regulates cyclin D1 in melanoma: Malme-3M
GSE18545	miR-9	A MicroRNA Expression Signature For Cervical Cancer Prognosis
GSE18625	miR-145	Identification of miR-145 targets involved in colon cancer
GSE18651	miR-29a,29b,29c	miR-29 targets in human fetal lung fibroblast IMR-90 cells
GSE19777	miR-221,222	Antisense miRNA-221/222 (si221/222) and control inhibitor (GFP) treated fulvestrant-resistant breast cancer cells
GSE20293	miR-30e,30e*	miR-30e* induced gene expression alteration in glioma cells
GSE20668	miR-100	Gene expression in human umbilical cord endothelial cells following premiR-100 overexpression
GSE20679	miR-517a	mRNA expression profile modified by microRNA mir-517a (MIR517A) in human hepatocellular carcinoma cell line
GSE20745	miR-17,18a,19a, 19b,20a,92a	Members of the microRNA-17-92 cluster exhibit a cell intrinsic anti-angiogenic function in endothelial cells
GSE21577	miR-103a,103b,93, 19b,106b	Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP: miRNA inhibition data
GSE24069	kshv-miR-K12-10a	miR-K10a expression and inhibition
GSE25224	miR-124	Persistence of seed-based activity following microRNA segmentation
GSE4246		Analysis of transcripts regulated by Dicer and Argonaut proteins in human HEK-293 cells (Ago2 knockdown)

3.2 CU-DREAM

CU-DREAM is physiogenomic discovery tool that compares two microarray experiments and determine if they share common regulated genes [6]. CU-DREAM is easy to use and compatible with GEO database. All materials about CU-

DREAM can be downloaded from http://pioneer.netserv.chula.ac.th/~achatcha/cu-dream.

We have performed intersection using CU-DREAM, between miRNA transfection experiment and AGO2 knockdown experiments, resulting in a 2x2 table as shown in Table 8.

		Ago2 knockdown experiment (GSE4246)	
		Down-regulated genes determined by Student's t-test	Other genes
miRNA transfection experiment	Up-regulated genes determined by Student's t-test	$\|A\|$	$\|B\|$
	Other genes	$\|C\|$	$\|D\|$

Table 8: Intersection between two microarray experiments through CU-DREAM.

3.3 Software

We wrote a piece of code to perform local sequence alignment using Smith-Waterman (SW) algorithm. The code is identical to the SW algorithm explained in Chapter 2.

CHAPTER IV

Experimental Results

The intersections between miRNA transfection experiments and Ago2 knockdown experiments are shown in Table 9 (sorted by odds ratio). If odds ratio > 1 and p-value is significant, we hypothesize that miRNAs could bind on promoter and activate transcription in gene set *A*. From the experimental Results we have applied only the top four results to local sequence alignments.

Table 9: The intersection results between miRNA transfection and Ago2 knockdown experiments.

GSE	2x2 table				Odds ratio	95% confidence interval	Unadjusted p-value (Pearson's Chi-square)
	$\|A\|$	$\|B\|$	$\|C\|$	$\|D\|$			
GSE19777 (1)	31	166	893	10,983	2.30	1.56 - 3.39	1.69E-05
GSE19777 (2)	26	151	898	10,998	2.11	1.38 - 3.21	3.90E-04
GSE20745	12	78	912	11,071	1.87	1.01 - 3.44	4.19E-02
GSE6207	31	214	893	10,935	1.77	1.21 - 2.60	2.94E-03
GSE24069	6	32	584	5,127	1.65	0.69 - 3.95	2.60E-01
GSE21577	50	383	869	10,650	1.60	1.18 - 2.17	2.14E-03
GSE13105	33	215	725	7,439	1.57	1.08 - 2.29	1.65E-02
GSE11701	19	128	706	6,837	1.44	0.88 - 2.34	1.43E-01
GSE18625	9	76	915	11,073	1.43	0.72 - 2.87	3.07E-01
GSE16571	70	545	790	8,544	1.39	1.07 - 1.80	1.26E-02
GSE16700	58	477	802	8,612	1.31	0.98 - 1.73	6.30E-02
GSE13105	35	276	723	7,377	1.29	0.90 - 1.85	1.59E-01
GSE13296	17	160	907	10,986	1.29	0.78 - 2.13	3.26E-01

GSE20668	56	435	789	7,918	1.29	0.97 - 1.72	8.02E-02
GSE16908	38	281	716	6,663	1.26	0.89 - 1.78	1.94E-01
GSE7864 (3)	71	566	722	7,257	1.26	0.97 - 1.63	7.81E-02
GSE7864 (2)	66	543	727	7,280	1.22	0.93 - 1.59	1.48E-01
GSE16568	29	256	831	8,833	1.20	0.81 - 1.78	3.51E-01
GSE16569	7	62	853	9,027	1.19	0.55 - 2.62	6.56E-01
GSE7864 (1)	60	520	733	7,303	1.15	0.87 - 1.52	3.25E-01
GSE7864 (7)	47	405	746	7,418	1.15	0.85 - 1.58	3.67E-01
GSE18510	64	560	781	7,793	1.14	0.87 - 1.49	3.38E-01
GSE7864 (4)	39	358	754	7,465	1.08	0.77 - 1.51	6.62E-01
GSE25224	45	433	703	7,249	1.07	0.78 - 1.47	6.68E-01
GSE7864 (8)	45	447	748	7,376	0.99	0.72 - 1.36	9.64E-01
GSE7864 (9)	33	346	760	7,477	0.94	0.65 - 1.35	7.32E-01
GSE20679	24	211	748	6,132	0.93	0.61 - 1.43	7.49E-01
GSE20293	60	640	754	7,350	0.91	0.69 - 1.20	5.21E-01
GSE18545	46	552	814	8,537	0.87	0.64 - 1.19	3.93E-01
GSE7864 (5)	32	400	761	7,423	0.78	0.54 - 1.13	1.85E-01
GSE18651	8	88	764	6,250	0.74	0.36 - 1.54	4.23E-01
GSE7864 (6)	27	407	766	7,416	0.64	0.43 - 0.95	2.74E-02
GSE16572	2	64	858	9,025	0.33	0.08 - 1.35	1.03E-01

As on experimental results, we have performed computer-based analysis through local sequence alignment to count the best alignments score between the transfected miRNAs and the promoters of the genes in set *A* are describe in the next page. We have collected miRNA sequences from mirBase public database. Our findings show at least 6 miRNAs (mir-221, mir-222, mir-124, mir-17, mir-20a, mir-92a) that could bind on promoter and activate transcription.

GSE19777_1 (miR-221)

```
miR-221   5'-agcuacau-ugucugcuggguuuc-3'     score: 33
             |||||||| :|||| ||  |||||
NM_002998 3'-tcgatgtatgcaga-ga-gcaaag-5'     antisense
             ↑
             at -299bp upstream from TSS

=======================================================

miR-221   3'-cuuugggucgu-cuguuacaucg-5'      score: 33
             ||||:|| ||| |:|: || |||
NM_022459 5'-gaaatcctgcacggcgctggagc-3'      sense
             ↑
             at -448bp upstream from TSS

=======================================================

miR-221   5'-agcua-cauugucugcuggguuuc-3'     score: 34
             |:|:| |||||||||| :::: ||
NM_152640 3'-ttggtggtaacagacgttttgcag-5'     antisense
             ↑
             at -540bp upstream from TSS

=======================================================
```

GSE19777_2 (miR-222)

```
miR-222    3'-ugggucauc-ggucuacaucga-5'      score: 29
              |||||| :|  | | || ||||
NM_021994  5'-acccagaggaaccgttgcagct-3'      sense
              ↑
              at -904 bp upstream from TSS

=======================================================

miR-222    5'-gcuacaucuggcuacugggu-3'         score: 29
              || |  |||| || |:||||
NM_002035  3'-cgttagagacggagggccca-5'         antisense
              ↑
              at -949 bp upstream from TSS

=======================================================

miR-222    5'-agcuacaucuggcuacugggu-3'        score: 29
              || ||||: ||||:| | :||
NM_015420  3'-tctatgtgtaccggt-agtca-5'        antisense
              ↑
              at -892 bp upstream from TSS

=======================================================
```

```
miR-222    5'-agcua--caucu-ggcuacugggu-3'    score: 29
              |||||  | ||| || | ||||:|
NM_000362  3'-tcgatcagcagatcctaggaccta-5'    antisense
              ↑
              at -200 bp upstream from TSS

=======================================================

miR-222    3'-ugg-gucaucgg-ucuaca--ucga-5'   score: 29
              ||| || |:||| | ||||  ||||
NM_007173  5'-accacaatggcctatatgtaaagct-3'   sense
              ↑
              at -900 bp upstream from TSS

=======================================================

miR-222    5'-gcuacaucuggcuacugggu-3'        score: 29
              ||:|| ||||| ||||| ::
NM_000060  3'-cggtggagacc-atgacgtg-5'        antisense
              ↑
              at -202 bp upstream from TSS

=======================================================

miR-222    5'-agcuacaucuggcuacugggu-3'       score: 29
              |||:| || | | ||||| ||
NM_006364  3'-tcggtttataac-atgacaca-5'       antisense
              ↑
              at -680 bp upstream from TSS

=======================================================

miR-222    3'-ugggucaucgguc-uacaucg-5'       score: 29
              :||||| ||::|| ||  |||
NM_018466  5'-gcccaggagttagaatcaagc-3'       sense
              ↑
              at -987 bp upstream from TSS

=======================================================

miR-222    5'-gcuacaucuggcuacugggu-3'        score: 29
              |  || |||| || |||||:
NM_018466  3'-ctttggagacagaggacccg-5'        antisense
              ↑
              at -702 bp upstream from TSS

=======================================================
```

```
miR-222    5'-gcuacaucuggcuacugggu-3'        score: 29
              | ||| :|||| : |||||:
NM_005596  3'-ctatggggacccgagacccg-5'        antisense
              ↑
              at -366 bp upstream from TSS

=====================================================

miR-222    5'-agcua--caucu-ggcuacugggu-3'    score: 29
              |||||  | ||| || | ||||:|
NM_000362  3'-tcgatcagcagatcctaggaccta-5’    antisense
              ↑
              at -200bp upstream from TSS

=====================================================

miR-222    5'-gcuacaucuggcuacugggu-3'        score: 29
              ||:|| ||||| ||||| ::
NM_000060  3'-cggtggagacc-atgacgtg-5'        antisense
              ↑
              at -202bp upstream from TSS

=====================================================
```

GSE20745 (miR-17, 20a, 92a)

```
miR-17     3'-gauggacgugacauucgugaaac-5'     score: 34
              :|:|:| :|:| |||||: ||||
NM_006644 5'-ttgcttttattataagcggtttg-3'      sense
              ↑
              at -658bp upstream from TSS

=====================================================

miR-20a    3'-gauggacgugauauucgugaaau-5'     score: 35
              :|:|:| :|:|||||||: |||:
NM_006644 5'-ttgcttttattataagcggtttg-3'      sense
              ↑
              at -658bp upstream from TSS

=====================================================

miR-92a    3'-ugucc-ggcccuguucacguuau-5'     score: 34
              | ||| |||||:|| |||||: |
NM_031372 5'-agaggcccggggcacgtgcagaa-3'      sense
              ↑
              at -169bp upstream from TSS

=====================================================
```

GSE6207 (miR-124)

```
miR-124   5'-uaaggcacgcggugaaugcc-3'         score: 30
             | |||| |||:|: |||| |
NM_005342 3'-actccgggcgtcggttaccg-5'         antisense
             ↑
             at -69bp upstream from TSS

=======================================================

miR-124   5'-uaaggcacgcgguga-augcc-3'        score: 30
             || :||||  :|||| |||||
NM_005107 3'-atctcgtgaatcactgtacgg-5'        antisense
             ↑
             at -648bp upstream from TSS

=======================================================
```

CHAPTER V

Discussion

We have proposed a number of gene promoters that could be targets of RNAa. Moreover, we have identified the putative binding locations where miRNAs could bind on promoter and trigger RNAa. There are several observations as follows.

- Unlike a synthetic RNA that is made complementary to a specific target, a perfect match between a natural miRNA and a promoter is very rare. It is expected that the binding quality determines on the effectiveness of RNAa. If so, the gene activation by a miRNA is modest compared to that of human-invented RNAs. This makes the identification of natural RNAa difficult because the difference of gene expression between cases (miRNA transfected) and controls is very small and it is hardly detected with few samples.

- Although the transfected miRNAs putatively bind on the promoters of genes in set *A*, they can also bind on the promoters of genes in set *D* as well. We use 2x2 tables below (see Table 10) to test whether the transfected miRNAs often bind on the set *A*, which is supposed to be the targets of RNAa. Unfortunately, the odds ratios are not high and the p-values are not significant. This suggests that there must be other necessary conditions for RNAa rather than RNA-DNA binding determined by sequence alignment.

Table 10: The association between alignment score and gene set *A*.

GSE19777 (miR-221)	|*A*|	|*D*|
Alignment score ≥ 33	3	1,267
Not bind	27	9,220

OR = 0.81, p-value = 7.27E-01

GSE20745	|*A*|	|*D*|
Alignment score ≥ 34	3	1334
Not bind	9	9236

OR = 1.38, p-value = 6.73E-01

GSE19777 (miR-222)	|*A*|	|*D*|
Alignment score ≥ 29	9	5,120
Not bind	17	5,377

OR = 0.56, p-value = 1.49E-01

GSE6207	|*A*|	|*D*|
Alignment score ≥ 30	2	320
Not bind	29	10,117

OR = 2.18, p-value = 2.76E-01

- We have also compared the bindings between set *A* and set *D* in order to find a binding characteristic that may distinguish RNAa. The characteristics of interest may be alignment score, sense/antisense, and distance from TSS. However, we have not yet found any remarkable characteristics. A large number of the bindings in set *D* are shown below.

GSE19777_1 (miR-221)

```
miR-221    3'-cuuugggucgucuguuacaucga-5'     score: 37
              ||||||:||||||:  || :|||
NM_032591  5'-gaaacctagcagattctggggct-3'         sense
              ↑
              at -652 bp upstream from TSS

=====================================================

miR-221    5'-agcuacauugucugcuggguuuc-3'     score: 37
              | || | ::|:||||||||||||
NM_005600  3'-tagaagcggcggacgacccaaag-5'     antisense
              ↑
              at -48 bp upstream from TSS

=====================================================
```

```
miR-221    5'-agcuacauugucugcuggguuu-3'      score: 36
              | || ||||||||||||: |:|
NM_001494  3'-tagaagtaacagacgactgaga-5'      antisense
              ↑
              at -859 bp upstream from TSS

==========================================================

miR-221    3'-ugggucgucuguuacaucga-5'        score: 35
              | |||||||::|:|||||||
NM_001755  5'-agccagcaggtagtgtagct-3'            sense
              ↑
              at -578 bp upstream from TSS

==========================================================

miR-221    5'-agcuacauugucugcuggguuuc-3'     score: 34
              ||||  |:| |||||  :|||||
NM_006659  3'-tcgacctgaaagacgtgtcaaag-5'     antisense
   ↑
   at -587 bp upstream from TSS

==========================================================

miR-221    5'-agcuacauugucugcug-gguuuc-3'    score: 33
              |||:||| :|:|||| :  |:|||
NM_138635  3'-tcggtgtcgcggacgttagcgaag-5'    antisense
              ↑
              at -315 bp upstream from TSS

==========================================================
```

GSE19777_2 (miR-222)

```
miR-222    5'-agcuacaucuggcuacugggu-3'       score: 37
              |||:|||| |||||||||:|:
NM_005789  3'-tcggtgtacaccgatgactcg-5'       antisense
              ↑
              at -976 bp upstream from TSS

==========================================================



miR-222    3'-ugggucaucggucuacaucga-5'       score: 36
              :|:|:||||||| ||||:|||
NM_030877  5'-gctcggtagccatatgtggct-3'       sense
              ↑
              at -365 bp upstream from TSS

==========================================================
```

```
miR-222    5'-gcuacaucuggcuacugggu-3'        score: 35
              ||:|| | ||||||||||||
NM_138330  3'-cggtgcacaccgatgaccca-5'        antisense
   ↑
   at -675 bp upstream from TSS

=========================================================

miR-222    3'-ugggucaucggucuacauc-ga-5'      score: 35
              |:||:|| |||:||||||| ||
NM_005803  5'-atccggttgccggatgtagtct-3'      sense
              ↑
              at -375 bp upstream from TSS

=========================================================

miR-222    5'-agcuacaucuggcuacugggu-3'        score: 34
              |||||||| ||| || |::||
NM_198450  3'-tcgatgtaaacccataattca-5'        antisense
              ↑
              at -328 bp upstream from TSS

=========================================================

miR-222    3'-ugggucaucggucuac-auc-ga-5'      score: 33
              |||:|  ||||||||| ||| ||
NM_019082  5'-acctaacagccagatgctagtct-3'      sense
              ↑
              at -808 bp upstream from TSS

=========================================================

miR-222    5'-agcuacaucug-gcua-cugggu-3'      score: 32
              ||||| |:||| |||| |||||:
NM_006717  3'-tcgat-tggactcgatggacccg-5'      antisense
              ↑
              at -589 bp upstream from TSS

=========================================================

miR-222    5'-gcuacaucuggcuacugggu-3'         score: 31
              || || ||:| || ||||||
NM_006590  3'-cgttggaggcggaggaccca-5'         antisense
              ↑
              at -453 bp upstream from TSS

=========================================================
```

```
miR-222     3'-ugggucaucggucuaca-ucga-5'      score: 30
               |||| | ||| ||:| | :|||
NM_018840   5'-accctggagcaaggtctgggct-3'      sense
   ↑
   at -337 bp upstream from TSS

=======================================================

miR-222     3'-gggucaucgg-ucuacaucga-5'       score: 29
              |||||||||: |||    |||
NM_016019   5'-cccagtagctgagaatacgct-3'       sense
              ↑
              at -678 bp upstream from TSS

=======================================================

miR-222         5'-cuacaucuggcuacugggu-3'     score: 29
                   || ||||| | |||||:::
NM_001099668    3'-gacgtagagcaatgacttg-5'     antisense
                   ↑
                   at -491 bp upstream from TSS

=======================================================
```

GSE20745(miR-17, miR-18, miR-19a, miR-19b, miR-20a, miR-92a)

```
miR-17      5'-caaagugcuuacagugcaggua-3'      score: 38
               | |||| ||| ||||||||||
NM_005776   3'-gattcaagaaagtcacgtccat-5'      antisense
               ↑
               at -256 bp upstream from TSS

=======================================================

miR-19a     3'-agucaaaac-guaucuaaacgugu-5'    score: 38
               ||||||||| :|||  ||||||:|
NM_006720   5'-tcagttttgttatacttttgcata-3'    sense
               ↑
               at -352 bp upstream from TSS

=======================================================

miR-19b     3'-agucaaaacguaccuaaacgugu-5'     score: 38
               ||||| |||:||| ||||| :||
NM_033641   5'-tcagtgttgtatgcatttgggca-3'     sense
               ↑
               at -312 bp upstream from TSS

=======================================================
```

```
miR-20a     5'-uaaagugcuuauagugcaggua-3'      score: 38
               |||||||||||  | |:||||:|
NM_020141   3'-atttcacgaatctaatgtccgt-5'      antisense
               ↑
               at -658 bp upstream from TSS

=======================================================

miR-92a     5'-uauugcacuuguccc-ggccugu-3'     score: 38
               ||:||| ||||||||  ||||||:
NM_021167   3'-atgacgagaacagggtccggacg-5'     antisense
               ↑
               at -993 bp upstream from TSS

=======================================================

miR-17      3'-auggacgugacauucgugaaac-5'       score: 37
               |:||| |||||||||:|:||||
NM_207293   5'-tgcct-cactgtaagtattttg-3'          sense
               ↑
               at -160 bp upstream from TSS

=======================================================

miR-18a     3'-gauagac-gugaucuacg-uggaau-5'    score: 37
               |||||||  ||  |||||||  |:  |||
NM_002230   5'-ctatctgccaatagatgcaatgtta-3'        sense
               ↑
               at -273 bp upstream from TSS

=======================================================

miR-19a     3'-agucaaaacguaucuaaacgugu-5'      score: 37
               |||||  |||:||:  |||||  :||
NM_033641   5'-tcagtgttgtatgcatttgggca-3'          sense
               ↑
               at -312 bp upstream from TSS

=======================================================

miR-19b       3'-agu-caaaacguaccua-aacgugu-5' score: 37
                    |||  ||  ||  |||||:|  |||||||
NM_001079812 5'-tcatgtgttccatgggtattgcaca-3'      sense
                  ↑
                  at -395 bp upstream from TSS

=======================================================
```

```
miR-20a    3'-gauggacgugauauucgugaaau-5'     score: 37
              || |||||||| |:| |||| ||
NM_206808  5'-cttcctgcactttgaccactcta-3'         sense
              ↑
              at -470 bp upstream from TSS

=====================================================

miR-92a    5'-auu-gcacuugucccggccugu-3'      score: 37
              ||| ||||:||:|||:||||||
NM_015015  3'-taaccgtggacggggtcggaca-5'      antisense
              ↑
              at -431 bp upstream from TSS

=====================================================

miR-17     5'-caaagugcuuacagugcagguag-3'     score: 36
              |||||||||   | |:||||:||
NM_007271  3'-gtttcacgaccttaatgtccgtc-5'     antisense
              ↑
              at -958 bp upstream from TSS

=====================================================
miR-18a    5'-ua-aggugcaucuagugc-agauag-3'   score: 36
              || |||:|||||| |: | ||||||
NM_144703  3'-atatccgcgtagaccgggttctatc-5'   antisense
              ↑
              at -515 bp upstream from TSS

=====================================================

miR-19a      3'-agu-caaaacguaucua-aacgugu-5' score: 36
                ||| || || |||:|:| |||||||
NM_001079812 5'-tcatgtgttccatgggtattgcaca-3' sense
                ↑
                at -395 bp upstream from TSS


=====================================================

miR-19b    3'-gucaaaacguaccuaaacgugu-5'      score: 36
              | ||| |||||| |||| ||||
NM_080655  5'-ctgttgtgcatgcatttacaca-3'      sense
              ↑
              at -576 bp upstream from TSS

=====================================================
```

```
miR-20a    3'-gauggacgugauauucg-ugaaau-5'    score: 36
              ||||:| ||||| | || |:||||
NM_173822  5'-ctacttccactacatgcaatttta-3'    sense
              ↑
              at -816 bp upstream from TSS

=====================================================

miR-92a    3'-uguccggcccuguu-cacguua-5'       score: 36
              ||||||| ||:||| |||||:|
NM_002162  5'-acaggcctgggcaaggtgcagt-3'       sense
              ↑
              at -77 bp upstream from TSS

=====================================================

miR-17     3'-gauggacgugacauuc-gugaaac-5'    score: 35
              |||: ||||: ||| | :||||||
NM_148899  5'-ctatatgcatggtatgttactttg-3'    sense
              ↑
              at -255 bp upstream from TSS

=====================================================

miR-18a    5'-uaaggugcaucuagugcagauag-3'     score: 35
              || :| :||||:|||:|| |:||
NM_003594  3'-atctcttgtaggtcatgtgtgtc-5'     antisense
              ↑
              at -880 bp upstream from TSS

miR-19a    3'-gucaaaacguaucuaaacgugu-5'      score: 35
              ||||||||||: ||  ||| |
NM_003813  5'-cagttttgcatgtatgagcaaa-3'      sense
              ↑
              at -816 bp upstream from TSS

=====================================================

miR-19b    5'-gugcaaauccaugcaa-aacuga-3'     score: 35
              | |||||| ||:|||| | ||||
NM_152551  3'-ctcgtttacgtgcgttctcgact-5'     antisense
              ↑
              at -132 bp upstream from TSS

=====================================================
```

```
miR-20a     3'-gauggacgugauau-ucgugaaau-5'     score: 35
               | ||:| ||:|| | ||||:||||
NM_032854   5'-ccacttacattaaacagcatttta-3'     sense
               ↑
               at -231 bp upstream from TSS

=====================================================

miR-92a     5'-uauu-gca-cuuguccc ggccugu-3'    score: 35
               ||:| ||| |:||:|||:||||:|
NM_003184   3'-atgatcgtgggacggggtcggata-5'     antisense
               ↑
               at -273 bp upstream from TSS

=====================================================

miR-17      3'-uggacgugacauucgugaaac-5'        score: 34
               :||||:| | | |||||||||
NM_001990   5'-gcctgtaatctcagcactttg-3'        sense
               ↑
               at -654 bp upstream from TSS

=====================================================

miR-18a     3'-gauagacgug-au-cu-acguggaau-5'   score: 34
               ||||:||||| || || ||||: | |
NM_14470    5'-ctatttgcacttatgagtgcatgtga-3'   sense
               ↑
               at -783 bp upstream from TSS

=====================================================

miR-19a     5'-ugugcaaa-ucuaugc-aaaacuga-3'    score: 34
               | || ||| ||| | | ||||||||
NM_153211   3'-aaacctttaagaaaggattttgact-5'    antisense
               ↑
               at -544 bp upstream from TSS

=====================================================
miR-19b       3'-gucaaaacguacc-uaaacgugu-5'    score: 34
                 | ||||| ||||| |||| || |
NM_001112706 5'-ctgtttttcatggtattttcaaa-3'     sense
                 ↑
                 at -576 bp upstream from TSS

=====================================================
```

```
miR-20a     5'-uaaagugcuuauagugcaggu-3'       score: 34
               |||||||||| ||| : |:||
NM_012287   3'-atttcacgaagatctttttca-5'       antisense
               ↑
               at -820 bp upstream from TSS

==========================================================

miR-92a       5'-uauugcacuugucccggccugu-3'    score: 34
                 ||||: ||||| |||||||:::
NM_001039083 3'-ataat-tgaaccgggccgggtg-5'     antisense
                 ↑
                 at -369 bp upstream from TSS

==========================================================
```

GSE6207 (miR-124)

```
miR-124     5'-uaaggcacgcggugaaugcc-3'        score: 33
               : |||||||||||||  |||
NM_003257   3'-gctccgtgcgccactgtcgg-5'        antisense
               ↑
               at -401 bp upstream from TSS

==========================================================

miR-124     3'-cguaaguggcgcacggaau-5'         score: 33
               |:|||| ||||| ||||||
NM_015914   5'-gtattccccgcgggcctta-3'         sense
               ↑
               at -332 bp upstream from TSS

==========================================================

miR-124     3'-ccguaaguggcgcacggaau-5'        score: 32
               || |||| || |||| ||||
NM_182767   5'-ggaattctccacgtggctta-3'        sense
               ↑
               at -723 bp upstream from TSS

==========================================================
```

```
miR-124    5'-aaggcacgcg-gugaaugc-3'        score: 31
              ||:||||||| | ||||||
NM_198321  3'-tttcgtgcgcactcttacg-5'        antisense
              ↑
              at -830 bp upstream from TSS

=========================================================

miR-124    5'-uaaggca-cgcggugaaugcc-3'       score: 30
              | |||||  ||| || |||||
NM_002873  3'-agtccgtcacgcgacctacgg-5'       antisense
              ↑
              at -981 bp upstream from TSS

=========================================================
```

This Research is published in International Conference on Bioinformatics and Biomedical Technology (ICBBT) at Singapore in 2012.

References

[1] A. Fire, C. Mello. **RNA Interference. Advanced Information on The Nobel Prize in Physiology 2006.**

[2] L. Li et al. **Small dsRNAs induce transcriptional activation in human cells.** PNAS 2006, 103 (46): 17337-17342.

[3] R. Place et al. **MicroRNA-373 induces expression of genes with complementary promoter sequences.** PNAS 2008, 105 (5): 1608-1603.

[4] R. Place et al. **Defining features and exploring chemical modifications to manipulate RNAa activity.** Curr. Pharm. Biotechnol. 2010, 11 (5): 518-526.

[5] T. Barrett et al. **NCBI GEO: archive for functional genomics data sets** – 10 years on. Nucleic. Acids. Res. 2011, 39 (Database issue): D1005-10.

[6] C. Aporntewan, A. Mutirangura. **Connection Up- and Down-Regulation Expression Analysis of Microarrays (CU-DREAM): A physiogenomic discovery tool, Asian Biomed.** 2011, 5 (2): 257-262.

[7] T. Smith, M. Waterman. **Identification of common molecular subsequences.** J. Mol. Biol. 1981, 147: 195–197.

[8] A. Kozomara, S. Griffiths-Jones. **mirBase: integrating microRNA annotation and deep-sequencing data.** Nucleic Acids Res. 2011, 39 (Database issue): D152-7.

[9] E. Check. **RNA interference: Hitting the on switch. Nature 2007**, 448: 855-858.

Biography

Mr. Nilesh Gramani was born in India. He obtained his degree in Bachelor of Commerce from the South Gujarat University, Gujarat, India in 2000.