Computational analysis of regulatory mechanism and interactions of microRNAs

Takaya Saito

Faculty of Medicine

Norwegian University of Science and Technology

A thesis submitted for the degree of

Philosophiae Doctor

2011

Abstract

For years, RNAs were thought to have only two broad functions in cells, transmitting information between DNA and protein as messenger RNA (mRNA), and playing structural, catalytic, information decoding roles in protein synthesis as ribosomal RNA (rRNA) and transfer RNA (tRNA). However, the discovery of RNA interference (RNAi) changed this picture. RNAi is a regulatory process that uses small non-coding RNAs (ncRNAs) to suppress gene expression at the post-transcriptional level. This discovery led to identification of many classes of functional ncRNAs. MicroRNA (miRNA) is a class of such ncRNAs with ∼22 nucleotides that are abundant and found in most eukaryotic cells. This thesis focuses on revealing regulatory roles and characteristics of miRNAs through bioinformatics approaches by addressing three research questions.

The first research question is whether we can enhance miRNAs target prediction in animals by considering multiple target sites. Many algorithms exist for miRNA target predictions, but most algorithms do not consider multiple target sites. Predicting accurate miRNA target genes is important to infer miRNA regulatory roles since annotations of miRNA regulations are still poor. To solve this possible fault, we developed a two step support vector machine (SVM) model. Benchmark tests showed that our two step model outperformed other existing miRNA target prediction algorithms.

The second research question is whether there are factors to explain differences between different miRNA high-throughput experiments. There are several high-throughput technologies widely used for miRNA experiments, such as microarray and quantitative proteomics, but the results from these technologies are often inconsistent. By statistically analyzing several such

high-throughput miRNA experiments, we revealed the characteristic of different technologies and also identified several factors that cause the differences.

The third research question is whether miRNAs interact with other classes of ncRNAs. There are strong evidences that some miRNAs are involved in transcription by interacting with other ncRNAs. We investigated ncRNAs in complex loci to find potential miRNA:ncRNA interactions. A complex locus is a locus that contains multiple genes that interact between themselves. We found evidence that some miRNAs are involved in transcriptional regulation with ncRNAs in complex loci.

In summary, this thesis provides solutions for these research questions, and it contributes to a better understanding of several important aspects of miRNA characteristics and regulations. It also shows effective bioinformatics approaches to develop a robust machine learning model and analyze different miRNA high-throughput experiments.

Acknowledgements

This thesis is based on four years of research funded by the Functional Genomics Program of the Norwegian Research Council. During the course of the research I have been helped by many individuals.

First and foremost, I would like to thank my two supervisors. I owe my deepest gratitude to Pål Sætrom for his contributions of time, ideas, and guidance to make this thesis possible. His wide knowledge and logical way of thinking have been of great value to me. I also gratefully acknowledge Finn Drabløs for his supervision. He has helped to make bioinformatics fun for me throughout my PhD.

I am indebted to the members of the Bioinformatics and Gene Regulation group for their contributions to providing an excellent working environment. I am especially grateful to Laurent Thomas and Even Skaland for their collaboration. I am also grateful to Tony Håndstad for his advice on manuscript preparation.

I would like to thank my friends and colleagues for making my time at NTNU tremendously enjoyable.

Lastly, I wish to thank my parents, Kikuo Saito and Mihoko Saito, for their love and support. Arigatou.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

List of Papers

Paper 1 **MicroRNAs - targeting and target prediction**
Takaya Saito and Pål Sætrom
New biotechnology 2010

Paper 2 **A two step site and mRNA-level model for predicting microRNA targets**
Takaya Saito and Pål Sætrom
BMC bioinformatics 2010

Paper 3 **Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments**
Takaya Saito and Pål Sætrom
(Submitted)

Paper 4 **MicroRNAs affect gene expression by targeting cis-transcribed non-coding RNAs**
Takaya Saito, Even Skaland, and Pål Sætrom
(Submitted)

Paper 5 **Inferring causative variants in microRNA target sites**
Laurent F. Thomas, Takaya Saito, and Pål Sætrom
Nucleic Acids Research 2011

Glossary

3′ UTR Three prime untranslated region; non-coding regions of mRNA on the 3′ end

5′ UTR Five prime untranslated region; non-coding regions of mRNA on the 5′ end

***k*-NN** *k*-nearest neighbor; a type of machine learning algorithm

C. elegans *Caenorhabditis elegans*; transparent roundworm about 1 mm in length

A Adenine; a purine nucleobase paired with thymine in DNA and uracil in RNA

ACC Accuracy; (TP + TN) / (P + N) in a binary classification model

ADTree Alternating decision tree; a machine learning algorithm that combines more than one decision tree

Agile Agile software development; a type of RAD methodology

Ago Argonaut protein; a key component of the RISC complex

ANN Artificial neural network; a machine learning method that mimics biological neural networks

ANOVA Analysis of variance; a statistical method to infer differences among multiple groups

AU rich Adenine:Uracil rich; nucleotide sequences with many adenines and uracils

AUC Area under the ROC curve; a performance measure to evaluate the ROC curves

bp Base pair; a unit for nucleotide length with a base pair as a Watson and Crick pair

C Cytosine; a pyrimidine nucleobase paired with guanine

CAR Chromatin associated RNA; experimentally validated non-coding RNAs that are associated with chromatin

cDNA Complementary DNA; DNA synthesized from mRNA by reverse transcriptase

CDS Coding sequence; coding region of mRNA

cis-NAT Cis-natural antisense transcript; a pair of sense and anti-sense transcript that overlap each other in the same locus

CLIP Cross-linking and immunoprecipitation; a technique used to pull down RNA-protein complexes

CNS Central nervous system; the central part of the nervous system in the brain

CROC concentrated ROC; a version of ROC for evaluating early retrieval performance

Cy3 Cyanine 3; a green fluorescent dye used in the microarray assay

Cy5 Cyanine 5; a red fluorescent dye used in the microarray assay

DGCR8 DiGeorge syndrome critical region gene 8; a protein that recognizes a miRNA stem loop in pri-miRNA

DNA Deoxyribonucleic acid; a nucleic acid that contains genetic information

dsRNA Double-stranded RNA; RNA with two complementary strands

EBI European bioinformatics institute; a center for research and services in bioinformatics in Europe

ERR Error rate; (FP + FN) / (P + N) in a binary classification model

EST Expressed sequence tag; a short subsequence of a cDNA sequence

FDR False discovery rate; FP / (FP + TN)

FLcDNA Full-length cDNA; full-length cDNA used by the Sanger sequencing method

FN False negative; prediction outcome is false while the actual value is true

FP False positive; prediction outcome is true while the actual value is false

G Guanine; a purine nucleobase paired with cytosine

G:U wobble Guanine:Uracil wobble; guanine and uracil wobble paring

Gb Giga base pair; 1,000,000,000 bp

GEO Gene expression omnibus; a public repository for microarray data

GFF General feature format; a text file format for genomic positional information

GPMDB Global Proteome Machine database; a public repository for proteomics data

GTP Guanosine triphosphate; a purine nucleotide that is used for energy transfer within the cell

HITS High throughput sequencing; the next generation sequencing

INSDC International nucleotide sequence database collaboration; a group that organizes SRA repositories

iTRAQ Isotope tags for relative and absolute quantification; a non-gel-based technique for quantifying proteins

K-S test Kolmogorov-Smirnov test; a non-parametric statistical method

LC-MS/MS Liquid chromatography-tandem mass spectrometry; MS/MS with liquid chromatography. Liquid chromatography separates ions or molecules dissolved in a solvent

MAF Multiple alignment format; a text file format for multiple alignments

MIAME Minimum information about a microarray experiment; a standard for reporting microarray experiments

miRISCs miRNA RISC; RISC loaded with miRNA

miRNA Micro RNA; a class of small ncRNA that regulates protein expression

ML Machine learning; a class of computational algorithms that can imitate learning

MS Mass-spectrometry; a technique that measures the mass-to-charge ratio of charged particles

MS/MS Tandem mass spectrometry; a technique that involves multiple steps of mass spectrometry

N Negative; actual negative values in a binary classification model

NB Naive Bayes; a type of statistical learning algorithm that uses Bayes' theorem

NCBI National center for biotechnology information; U.S. government-funded national resource for molecular biology information

ncRNA Non-protein-coding RNA; functional RNA that is not translated into protein

NPV Negative predictive value; TN / (TN + FN) in a binary classification model

OOP Object-oriented programming; a computer programming paradigm

P Positive; actual positive values in a binary classification model

PCR Polymerase chain reaction; a technique used to amplify DNA sequences

piRNA Piwi-interacting RNA; siRNA/miRNA like ncRNAs found in germline cells

Pol II RNA polymerase II; an enzyme that synthesizes several types of RNAs

Pol III RNA polymerase III; an enzyme that synthesizes rRNA, tRNA and other small RNAs

PPV Positive predictive value; TP / (TP + FP) in a binary classification model

PRC Precision; equivalent to PPV or Positive predictive value

pre-miRNA precursor miRNA; miRNA precursor with a hairpin stem loop, that is exported into cytoplasm

pri-miRNA primary miRNA; a RNA molecule that contains one or more miRNA stem loops

PRIDE Proteomics identifications database; a public repository for proteomics data

QP Quadratic programming; A class of optimization algorithms to maximize a quadratic function subject to linear constrains

RAD Rapid application development; a software development methodology

Ran Ras-related nuclear protein; a GTP binding protein that is involved in transport between nucleus and cytoplasm

RBF Radial basis function; a real-valued function whose value depends only on the distance from the origin

RDB Relational database; a computational data storage method. Data are stored in tables with a collection of relations

RIP Ribonucleoprotein immunoprecipitation; a technique used to pull down RNA-protein complexes

RISC RNA-induced silencing complex; a key multiprotein complex in RNAi

RITS RNA-induced initiation of transcriptional gene-silencing; a complex involved in regulation of chromatin structure

RNA Ribonucleic acid; a nucleic acid that catalyzes with many biological molecules

RNAi RNA interference; a regulatory process that suppresses gene expression at the post-transcriptional level with small RNAs

RNase Ribonuclease; an enzyme that degrades RNAs into smaller components

ROC Receiver operating characteristics; a graph that shows true positive rate versus false positive rate

rRNA Ribosomal RNA; RNA components of the ribosome

RT-qPCR Reverse transcription quantitative PCR; a variant of PCR that can be

used to measure RNA expression levels

SAGE Serial Analysis of Gene Expression; a sequencing technique that uses short tags generated from 3′ ends of mRNA transcripts

SILAC Stable isotope labeling with amino acids in cell culture; a technique for in vivo incorporation of a label into proteins

siRISC siRNA RISC; RISC loaded with siRNA

siRNA Small interfering RNA; small ncRNAs involved in RNAi for gene silencing

SN Sensitivity; TP / P in a binary classification model

SNP Single-nucleotide polymorphism; DNA polymorphism with a single nucleotide difference between members of a species

SP Specificity; TN / N in a binary classification model

SQL Structured query language; a language used with RDB

SRA Sequence read archive; data repository for next generation sequencing data

SRM Structural Risk Minimization; a machine learning principle

ssRNA Single-stranded RNA; RNA with one strand

SVM Support vector machine; a machine learning algorithm that guarantees the maximum margin between decision boundaries

SVR Support vector regression; a version of SVM for regression

TDD Test-driven development; a type of RAD methodology

TN True negative; prediction outcome is false while the actual value is false

TNR True negative rate; equivalent to SP or Specificity

TP True positive; prediction outcome is true while the actual value is true

TPR True positive rate; equivalent to SN or Sensitivity

tRNA Transfer RNA; transfer a specific amino acid for protein synthesis

U Uracil; a pyrimidine nucleobase paired with adenine in RNA

UV Ultraviolet; electromagnetic radiation with shorter wavelength than visible light

VC dimension Vapnik Chervonenkis dimension; a measure of capacity for the data point separation by hyperplanes

WTSS Whole transcriptome shotgun sequencing; high throughput technique at the whole transcriptome level with next generation sequencing

XP Extreme programming; a type of RAD methodology

1

Introduction

1.1 Challenges

In the 1980s, most non-protein-coding regions in the genome were thought to be 'junk' DNA with no functional purpose (1). Nonetheless, during the last two decades, new classes of non-coding RNAs (ncRNAs) that have gene regulatory roles have been discovered within these 'junk' regions (2). MicroRNAs (miRNAs) are one such new class of ncRNAs that have many important regulatory roles on a genome-wide scale (3). Because of their importance, research on miRNAs has gained popularity in recent years (Fig. 1.1). Although many research efforts have revealed basic miRNA characteristic and regulation (3, 4), there are still many challenges to identify the comprehensive characteristics and precise regulatory mechanism of miRNAs. This thesis covers three such challenges related to miRNA studies.

The first challenge is to identify accurate miRNA targets in animals. It is important to understand miRNA contributions to the genome-wide gene regulation, but there are several known obstacles related to indentifying miRNA targets in animals. Firstly, miRNAs are quite abundant (3), and one miRNA can potentially regulate many protein-coding genes (5). In some cases, miRNAs bind their target mRNAs by base-pairing with only six nucleotides (5), which results in thousands of potential candidate genes influenced by one miRNA at a genome-wide level. Secondly, since miRNAs are expressed in a cell- or tissue-specific manner (6), one true positive miRNA target can be a false positive in a different cell or tissue type. Thirdly, the precise mechanism of miRNA binding process on its target mRNA is unknown (4). Therefore, combinations

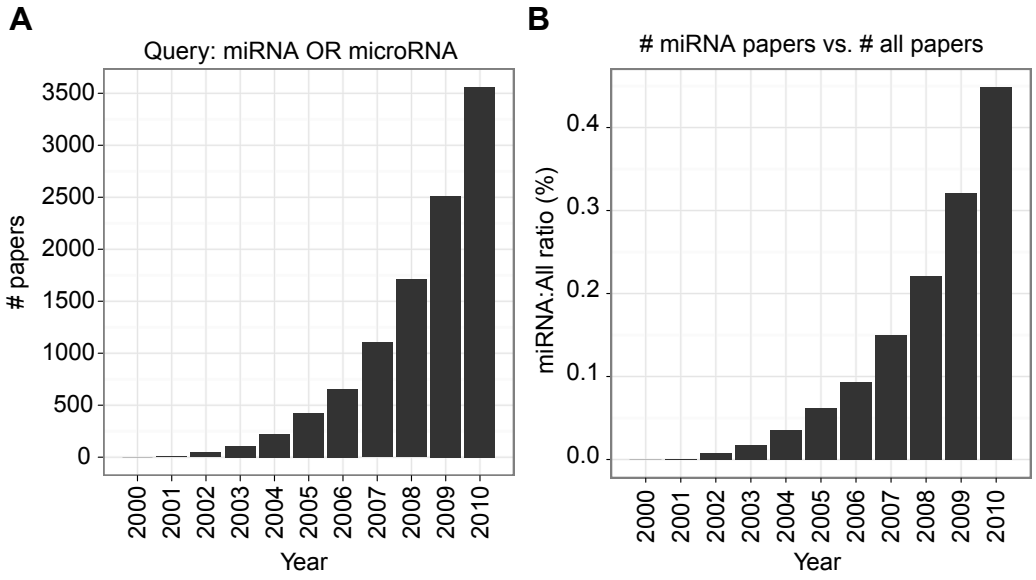

Figure 1.1: PubMed query - Two figures show the trend of miRNA related papers as (A) the number of papers, and (B) the ratio to all papers in PubMed.

of miRNA features are usually used to predict miRNA targets, but the combined effects of these features on miRNA targeting are unclear.

The second challenge is to interpret miRNA high-throughput data appropriately with high accuracy. Microarray, next generation sequencing, and quantitative proteomics are three major high-throughput technologies widely used for miRNA studies. Nonetheless, analyses of the data from these high-throughput technologies often give different interpretations regarding miRNA characteristics and regulation (7, 8, 9). A major obstacle is that there are many factors involved in these analyses, but the main factors that cause these differences are unknown.

The third challenge is to indentify potential miRNA interactions with other ncRNAs. Although most miRNAs regulate genes at the post-transcriptional level, some miRNAs can also regulate transcription itself (10, 11, 12). This transcriptional regulation seems to involve ncRNAs overlapping or interacting with the target gene promoters (13, 14, 15, 16, 17). Many aspects of this miRNA regulation at the transcription level are poorly understood. Moreover, few experimental data are available for this miRNA regulation at the transcription level.

1.2 Goals

The main goal of this thesis is to reveal the characteristics and regulations of miRNAs by analyzing several different types of high-throughput data through bioinformatics approaches. To achieve this goal, I defined three sub-goals to solve the three challenges of miRNA studies described in the previous section.

The first sub-goal is to develop a miRNA target prediction algorithm with high accuracy. Most existing prediction algorithms focus on identifying individual target sites without considering multiple target sites. They do not include multiple target sites that possibly contribute to miRNA regulation. Moreover, most algorithms use strict filtering, such as filtering with evolutional conservation. Filtering can reduce false positive miRNA targets, but it potentially removes many true positive targets at the same time. Therefore, the aim of this sub-goal is to develop a model that can predict unbiased miRNA targets by considering multiple targets without filtering.

The second sub-goal is to analyze several different types of miRNA high-throughput technologies. The aim of this sub-goal is to reveal the characteristics of each technology and identify strong factors that cause inconsistent results between different types of experiments by statistical approaches.

The third sub-goal is to infer potential miRNA regulations outside of $3'$ untranslated regions (UTRs) in general and interactions between miRNAs and ncRNAs in complex loci in particular. A complex locus is a region of DNA that contains multiple genes that have interactions between them or share common regulatory mechanisms (18). Our hypothesis is that some miRNAs interact with ncRNA:mRNA pairs in complex loci. The aim of this sub-goal is to investigate this hypothesis of miRNA involvement in complex loci together with miRNA regulations outside of $3'$ UTRs by computationally analyzing the data from high-throughput experiments.

In this thesis, these sub-goals are referred to in italic to clarify the relationship between parts of the text and their corresponding sub-goals if necessary.

- First sub-goal: *miRNA target prediction*
- Second sub-goal: *miRNA high-throughput experiments*
- Third sub-goal: *miRNA and other ncRNAs*

1.3 Thesis structure

This thesis consists of eight chapters followed by five papers.

Chapter Two: Papers and their corresponding sub-goals. This chapter summarizes the five papers included in this thesis. It also relates them to each sub-goal.

Chapter Three: MicroRNAs and other non-coding RNAs. This chapter introduces the history, characteristics, and biological functions of miRNAs as well as some additional information about other ncRNAs.

Chapter Four: High-throughput biological experiments. This chapter focuses on three high-throughput technologies used in our research: microarray, next generation sequencing, and quantitative proteomics.

Chapter Five: Statistical tests and methods. This chapter starts with explaining basic statistical tests followed by applied statistical approaches used throughout in our research, such as non-parametric tests, resampling, and multiple comparison tests.

Chapter Six: Machine learning theory and Support vector machine. Support vector machine (SVM) is the main method used in the first sub-goal: *miRNA target prediction*. This chapter explains the theoretical background of SVM, data preparation and evaluation methods for SVM, as well as some other machine learning methods for comparison.

Chapter Seven: Computational implementation. Any state-of-the-art model or algorithm is ineffective without appropriate computational implementation. This chapter focuses on the computation implementations used in our research.

Chapter Eight: Future perspective. This chapter describes potential improvements of each sub-goal as future perspectives.

2

Papers and their corresponding sub-goals

This thesis includes five papers, and each paper has a corresponding sub-goal (Table 2.1). This chapter gives a brief description of paper in context of each sub-goal.

Table 2.1: Papers and corresponding sub-goals of our research

Sub-goal	**Paper**	
miRNA target prediction	Paper 1	MicroRNAs - targeting and target prediction
	Paper 2	A two step site and mRNA-level model for predicting microRNA targets
	Paper 5	Inferring causative variants in microRNA target sites
miRNA high-throughput experiments	Paper 3	Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments
miRNA and other ncRNAs	Paper 4	MicroRNAs affect gene expression by targeting cis-transcribed non-coding RNAs

2.1 Three papers for the first sub-goal: *miRNA target prediction*

Paper 1: MicroRNAs - targeting and target prediction. This review paper outlines the features associated with animal miRNA targeting. It summarizes the characteristics of the features in six different categories: miRNA:mRNA paring, Site location, Conservation, Site accessibility, Multiple sites, and Expression profiles. It also contains a list of 30 different miRNA target prediction tools with information of feature coverage in context of the six categories.

Paper 2: A two step site and mRNA-level model for predicting microRNA targets. This paper presents a miRNA target prediction model that recognizes both the individual characteristics of functional binding sites and the global characteristics of miRNA-targeted mRNAs. Our novel two-step SVM model trains site level features at the first step, and, subsequently, it trains mRNA level features at the second step. Benchmark experiments showed that our two-step SVM model had a higher overall performance than other established miRNA target prediction tools.

Paper 5: Inferring causative variants in microRNA target sites. This paper shows an example that miRNA predictions from our two-step SVM model performs better than the other prediction algorithms when the predictions are used by other tools. Laurent F. Thomas was the main contributor to this study, and he developed a tool that can help identifying Single-nucleotide polymorphisms (SNPs) associated with diseases by focusing on SNPs affecting miRNA regulation. The tool uses miRNA target predictions to check the influence of SNPs that affect miRNA targeting. It can use any miRNA prediction tools that generate scores of miRNA target predictions. The paper showed that the tool had the best performance when our two-step SVM model was used.

2.2 One paper for the second sub-goal: *miRNA high-throughput experiments*

Paper 3: Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments. This paper shows characteristics of different miRNA

high-throughput experiments. Analysis on these high-throughput experiment data sometimes show inconsistent miRNA regulation factors, for example, one experiment shows 3′ UTR length is one of the most important factors, whereas other experiment shows it is least important. We investigated several factors that might affect this inconsistence, and we revealed that competition between endogenous miRNAs and the ectopically expressed miRNAs significantly contributed to the differences among different miRNA high-throughput experiments. We also found that this competition effect affected other factors, such as mRNA expression level and 3′ UTR length, in terms of miRNA targeting.

2.3 One paper for the third sub-goal: *miRNA and other ncRNAs*

Paper 4: MicroRNAs affect gene expression by targeting cis-transcribed non-coding RNAs. This paper shows potential miRNA regulation on two types of complex loci: cis-natural antisense transcripts (cis-NATs) and chromatin associated RNAs (CARs). We used several different types of data from high-throughput miRNA experiments to infer potential miRNA regulation on such loci. Our statistical analyses revealed that complex loci containing non-coding cis-NATs or CARs appeared to be under strong regulation, although this type of miRNA targeting is less prevalent than miRNA targeting of 3′ UTRs.

2. PAPERS AND THEIR CORRESPONDING SUB-GOALS

3

MicroRNAs and other non-coding RNAs

Since our main goal is to reveal the miRNA regulation and characteristics, this chapter introduces several different aspects of miRNAs, such as the history of miRNA discovery, miRNA biogenesis, and mechanism of miRNA regulation. In addition to miRNAs, it also describes several other classes of ncRNAs and their potential interactions with miRNAs. Specifically, analysis on such ncRNAs is the main objective of the third sub-goal: *miRNA and other ncRNAs*.

3.1 Thousands of miRNAs have been identified since the first miRNA discovery of *lin-4* in 1993

In 1993, Lee et al. found that *lin-4*, a gene involved in development timing in *C. elegans*, produces a small ncRNA instead of a messenger RNA (mRNA) (19). *lin-4* was known to regulate *lin-14*, but the protein product of *lin-4* had been undetected. The result of an alignment analysis indicated that *lin-4* has multiple complementary sites on the $3'$ UTR of *lin-14* (19). Further experiment revealed that this small ncRNA produced by *lin-4* can directly suppress the expression of *lin-14* by base-paring on the $3'$ UTR. This was the first discovery of this functional ncRNA with about 22 nucleotides that regulates specific protein expression by base-paring on the $3'$ UTR of its target mRNA (3). However, there were no other *lin-4*-like small ncRNAs identified, and this peculiar regulation of *lin-4* was recognized as a rare case (3).

Meanwhile, Fire et al. reported that they observed a gene silencing effect after injecting double-stranded RNAs (dsRNAs) into *C. elegans* (20) in 1998. They coined the term, RNA interference (RNAi), to describe this gene-silencing mechanism. Soon thereafter, RNAi was found to silence genes at the post-transcriptional level with small RNA molecules with 20-25 nucleotides, called small interfering RNAs (siRNAs) (21, 22, 23).

Then, nearly seven years after the discovery of *lin-4*, Reinhart et al. identified a *lin-4*-like small ncRNA, *let-7*, in *C. elegans* in 2000 (24). They revealed that *let-7* is also an ncRNA with about 22 nucleotides that regulates specific protein expression by base-paring on the $3'$ UTR. Since *let-7* has homologs in various species, this discovery led to identification of many other *let-7*- and *lin-4*-like small ncRNAs in other animals, including human and *Drosophila* (25). The term microRNA (miRNA) was coined to refer to these small ncRNAs of about 22 nucleotide length (26, 27, 28). Moreover, like siRNAs, miRNAs appeared to use the RNAi pathway to regulated genes at the post-transcriptional level (3).

Today, miRNAs are recognized as a very common class of ncRNAs that can regulate protein expression (3, 26, 27, 28). For instance, miRBase (29, 30, 31, 32), which is the main database for miRNA annotations, contains 15172 entries in 142 species as of release 16, 2010. They are abundant and found mostly in eukaryotes as well as in some viruses. MicroRNAs are known to play many important regulatory roles in eukaryotes (3), whereas some viruses encode viral miRNA genes that potentially regulate fundamental cellular processes both in the viruses and in their host cells (33, 34).

3.2 MicroRNA biogenesis involves multiple steps

Although the precise mechanism of miRNA biogenesis is unknown, Figure 3.1 shows the most widely accepted view of the miRNA biogenesis to date. The biogenesis involves multiple processes both in the nucleus and the cytoplasm.

First, RNA polymerase II (Pol II) transcribes a miRNA gene on the chromosome from DNA to single-stranded RNA (ssRNA) with $5'$ cap and poly-A tail (35). This ssRNA called primary miRNA (pri-miRNA) can be several hundreds or thousands nucleotides long, and it may contain one or more hairpin loops (36). An enzyme called DiGeorge Syndrome Critical Region 8 (DGCR8) recognizes the hairpin loop

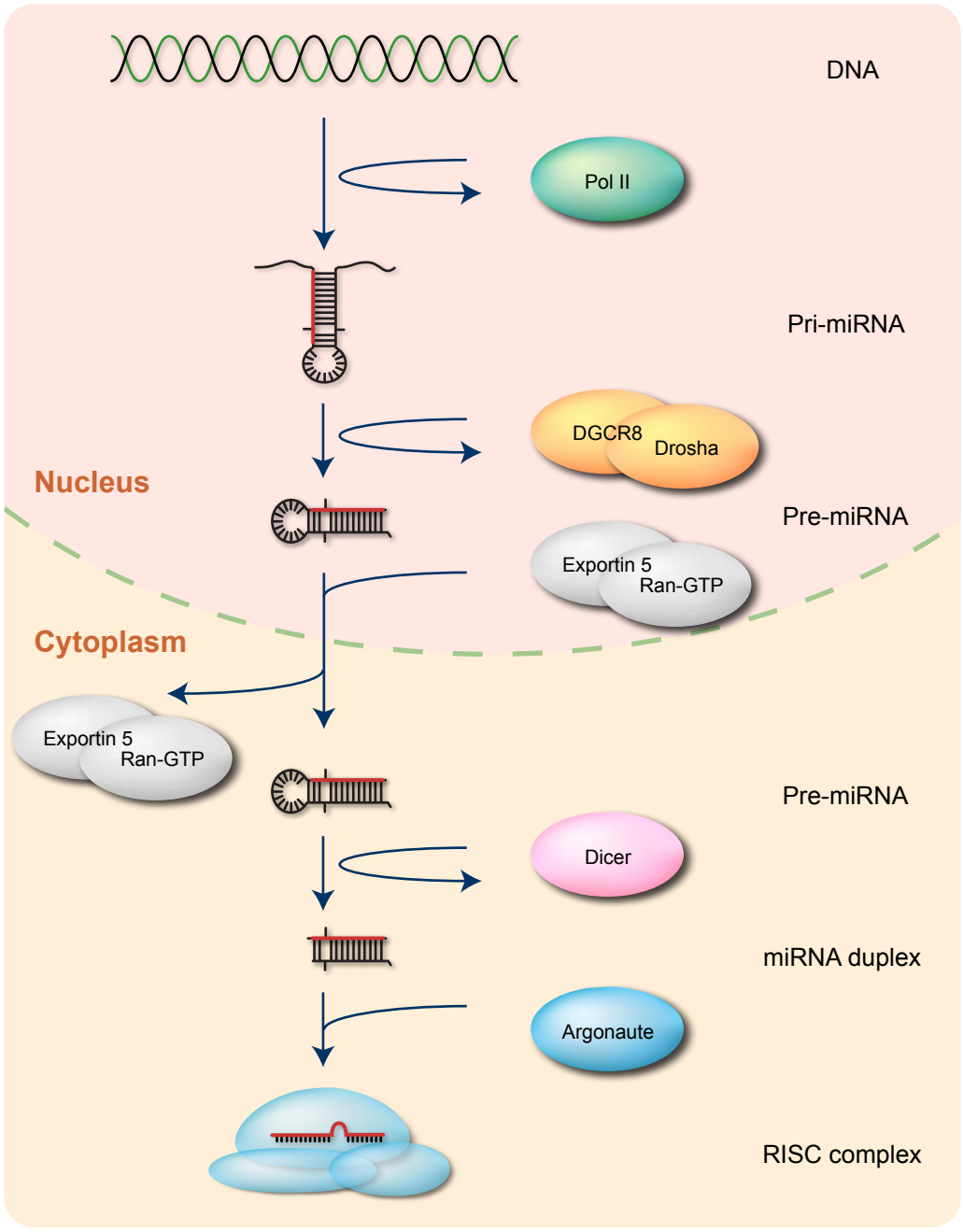

Figure 3.1: MicroRNA biogenesis - The figure shows the overview of miRNA biogenesis. Pol II transcribes ssRNA from DNA. The ssRNA forms pri-miRNA, that is further processed to a hairpin loop structure called pre-miRNA by DGCR8 and Drosha. Exportin 5 together with Ran-GTP exports pre-miRNA into the cytoplasm. Dicer cleaves pre-miRNA to form miRNA duplex. Only one strand of the miRNA duplex is usually bound to the Argonaute protein and loaded into the RISC complex.

in pri-miRNA, and a DGCR8 associated enzyme, called Drosha, cleaves the hairpin from the pri-miRNA (37, 38, 39, 40, 41). This cleavage results in a hairpin structure with approximately 60 nucleotides called precursor miRNA (pre-miRNA) (37, 42, 43, 44). Exportin 5 together with Ran-GTP exports pre-miRNAs from the nucleus to the cytoplasm (45, 46).

In the cytoplasm, a Ribonuclease (RNase) III enzyme, called Dicer, cleaves the loop of pre-miRNA and produces a miRNA:miRNA* duplex with approximately 22 nucleotides (37). Only one strand of this duplex usually becomes a mature miRNA as a guide to the target mRNA, and the other strand, defined as miRNA star-strand or miRNA* (27), is eventually degraded (3). The mechanism of the strand selection is unclear, but a strand that is less thermodynamically stable at its $5'$ end appears to be favored in some cases (47, 48). Argonaute (Ago) proteins are key proteins for miRNA targeting (49). Ago2, which is one of the Ago clade proteins, is mainly associated with mature miRNAs in mammals (50). A protein complex called RNA-Induced Silencing Complex (RISC) (51) that incorporates Ago2, uses the mature miRNA as a guide to bind and then catalyze specific target mRNAs (50).

Moreover, there are several known alternative pathways for miRNA biogenesis. For instance, some intronic miRNAs bypass Drosha processing by directly forming a pre-miRNA-like hairpin structure. These intronic miRNAs are defined as mitrons (52, 53) first identified in *Drosophila* and *C. elegans* (53), and also found in mammals (54). Another example of alternative miRNA pathways is that RNA polymerase III (Pol III) instead of Pol II transcribes some miRNAs (3), especially those residing upstream of Alu sequences (55). Alu sequences or Alu elements are abundant mobile elements especially found in the primate genomes (56).

3.3 RNA interference is the central mechanism for gene regulation by miRNAs and small interfering RNA

MicroRNAs and siRNA regulate and control gene expression through RNAi (23, 50). Although siRNAs have biochemically indistinguishable mature forms of ssRNAs from those of miRNAs, the siRNA biogenesis pathway to its mature form is different from that of miRNA (3). Dicer processes siRNA precursors, such as long dsRNAs or small

hairpin RNAs (shRNAs), and cleaves them into a ∼22 nt dsRNA with 2-nt 3′ overhangs (3, 50, 57). This dsRNA form of siRNA is essentially equivalent with the miRNA:miRNA* duplex.

Major functionalities of siRNAs and miRNAs are that siRNAs are defenders against foreign or invasive nucleic acid molecules such as viruses, transposons, and transgenes (23), whereas miRNAs are regulators of endogenous protein-coding genes (50). Moreover, exogenous siRNAs are widely used in gene knockdown experiments, and they are potentially useful for gene therapy. Exogenous siRNAs are directly introduced into the cytoplasm or taken up from the environment (23, 50).

The central mechanism of RNAi is that si/miRNAs loaded in RISC act like guides to bind the sites of their target mRNAs. The RISCs loaded with miRNAs are called miRISCs, whereas the RISCs with siRNAs are called siRISCs (50). RNAi has several different gene silencing modes (Fig. 3.2). The common gene silencing mode for siRNAs and plant miRNAs is cleaving mRNAs that have near-perfect complementary sites with the si/miRNAs (Fig. 3.2) (3). Two known gene silencing modes for animal miRNAs are transcriptional repression and mRNA degradation (Fig. 3.2). In either mode, miRNAs mainly target mRNAs that have partial complementary sites on their 3′ UTRs (3). Although translational repression was initially thought as the major regulatory model of animal miRNAs, a recent study used ribosome profiling assay and reported that most target genes of animal miRNAs were actually degraded (58). The precise mechanism of this degradation is still unclear, but it is possibly associated with deadenylation, decapping, and exonucleolytic digestion of the mRNA (59, 60, 61).

Moreover, exogenous siRNAs are known to act like miRNAs and down-regulate numerous unintended mRNAs in the same way as miRNA gene silencing. This unintended effect is called siRNA off-targeting (62). Considering this siRNA off-targeting is very important to design effective exogenous siRNAs.

3.4 MicroRNAs have various regulatory roles that are associated with important physiological and pathological processes

Many miRNAs play important regulatory roles by negatively controlling the expression level of mRNAs (3), and current estimates indicate that at least 60% of human protein-

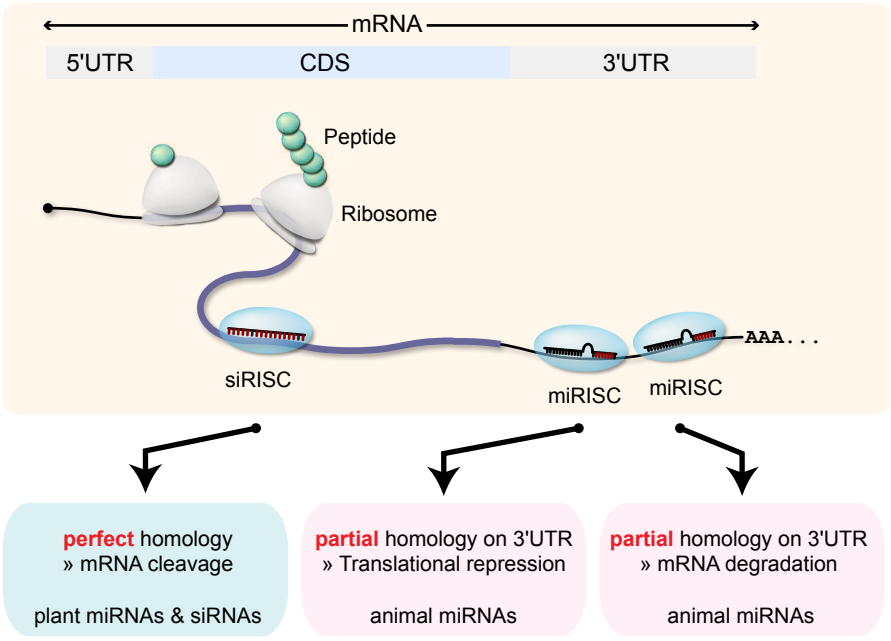

Figure 3.2: miRNA target - The figure shows three examples of RNAi regulation by miRISC and siRISC. The long curved line represents mRNA that is separated into three regions, 5′ UTR, CDS, and 3′ UTR. Ribosomes synthesize peptides while moving through the mRNA from 5′ to 3′ direction. One siRISC binds on the CDS and two miRISCs bind on the 3′ UTR of mRNA, in this example. Plant miRNA and siRNA have nearly perfect complementary which results in mRNA cleavage. Animal miRNAs require only partial complementary, and they either repress translation or contribute to mRNA degradation.

coding genes are under some influence of miRNAs (5). Many miRNAs, like *lin-4* and *let-7*, are involved in cell development processes (63). Other experimentally validated miRNA regulatory roles can be found in many cellular processes, such as growth control, differentiation, stem cell and germline proliferation, and apoptosis (64). However, annotations of many miRNA regulations are still poor, therefore, predicting accurate miRNA target genes is important to infer miRNA regulatory roles.

MicroRNAs are also associated with human diseases because of their wide range of gene regulatory roles (64). Several studies revealed miRNA involved diseases, such as cancer (65, 66), heart disease (67, 68), DiGeorge syndrome (64), Alzheimer's disease, and central nerve system (CNS) disorders (69). Moreover, some viruses encode viral miRNA genes (33). These viral miRNAs potentially regulate fundamental cellular processes both in the viruses and in their host cells (34).

3.5 Multiple properties of miRNA target recognition may enhance target efficacy in animal

Since most animal miRNAs have only partial complementary to their target mRNAs (4), miRNA targeting usually requires additional features for better target recognition. The most important feature is the seed type, which is the region for the partial complementary. The seed site contains six nucleotides from position 2 to 7 of the miRNA (4, 70, 71), as the position begins with 1 at the $5'$ end of the miRNA. Even though the definition of the seed site is ubiquitous, the definition of the seed types is different among different studies. Figure 3.3 shows nine common seed types that are widely accepted in many studies. Seed types consist of stringent and non-stringent groups. Three seed types, 7mer-A1, 7mer-m8, and 8mer, belong to the stringent group (Fig. 3.3). These seed types have perfect Watson-Crick paring in their seed sites, and they are usually stronger than those in the non-stringent groups in terms of miRNA target recognition (4). 7mer-A1 has an adenine (A) at position 1 of the target mRNA. An adenine at position 1 of the mRNA is known to enhance miRNA target recognition (72). 7mer-m8 has paring at position 8. 8mer has an adenine at position 1 and paring at position 8. The non-stringent group consists of 6mer, two G:U wobble, one loop, and two bulge types (Fig. 3.3) (73). They are less effective than the stringent group, but they are still functional since miRISC can tolerate some mismatches (74). 6mer

has perfect seed paring, whereas the other types are equivalent to 8mers except one mismatch or wobble paring. LP has a loop in the seed site. GUM has a G:U wobble site on the miRNA whereas GUT has a wobble site on the target mRNA. Similarly, BM has a bulge on the miRNA whereas BT has a bulge on the target mRNA.

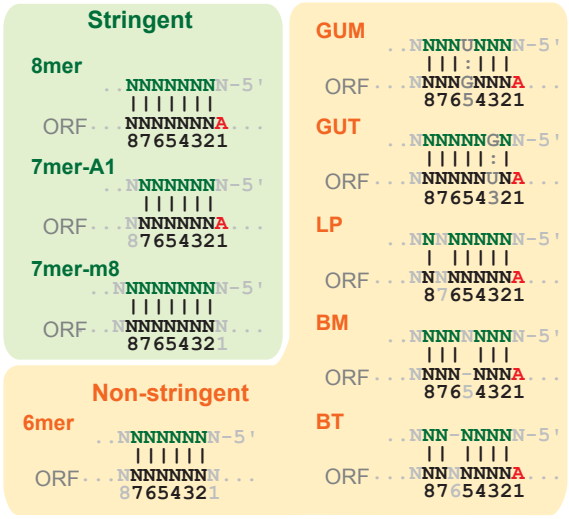

Figure 3.3: miRNA seed types - Examples of three stringent seed types (8mer, 7mer-A1, and 7mer-m8), and six non-stringent seed types (6mer, GUM, GUT, LP, BM, and MT). The strand on the top of each seed type represents position 1-8/9 of miRNA, and the bottom strand represents target mRNAs.

Additional paring at the 3′ part of miRNAs can increase the efficacy of miRNA repression, and it can also compensate for a seed mismatch to create a functional site (4). Three to four matches at position 13-16 for stringent seeds, and four to five matches at position 13-19 for non-stringent seeds are known as 3′ supplementary and 3′ compensatory paring, respectively (4, 75).

Many target sites are well conserved among closely related species (5). However, there are many approaches to define "well conserved" targets. For instance, some use perfect seed matches among several species (76, 77, 78), whereas others use pre-defined conservation scores calculated by global phylogenetic analysis (79, 80). Moreover, even though many targets are well conserved, some targets are also species-specific. For instance, one study predicted that about 30% of all experimentally validated targets are poorly conserved (81).

3.5 Multiple properties of miRNA target recognition may enhance target efficacy in animal

The site accessibility of miRISC can be measured by computing the secondary structure of target sites through minimum free energy approaches. The site accessibility is potentially a very strong feature to predict true miRNA target sites because it is directly linked to target recognition. However, the precise mechanism of miRISC access on its targets is unknown, and developing a precise prediction model with mRNA secondary structure calculation usually requires huge computational power. Some models used elaborate two step approaches with the first step as initial forming of miRNA:mRNA complex, and the second step as hybrid elongation to form the complete miRNA:mRNA complex (80, 82). An alternative feature to site accessibility is to measure the occurrence of AU rich elements both upstream and downstream of the seed site (75). Thermodynamic stabilities of AU base pairs are much lower than that of GC base pairs, and it can be a reason that sites surrounded by AU-rich context has better site accessibility for miRISCs. Moreover, this AU rich approach is computationally inexpensive with a good prediction performance (75). Scoring AU context is a more reliable measurement of target accessibility than any other currently available models with mRNA secondary structure calculation (4).

Multiple target sites enhance miRNA target efficacy (83). Although the general effect of multiple sites is additive, the effect can be synergistic when two miRNA targets are within optimal distance. One reported such optimal distance is defined as two seed sites separated by between 13 and 35 nt (84). For example, a gene with two 7mers within optimal distance is more down-regulated than a gene with a single 8mer. However, the effect is not apparent when two 7mers are not within optimal distance. In this case, a gene with 8mer is more down-regulated than a gene with two 7mers (4).

To summarize miRNA target recognition and efficacy, most animal miRNAs bind on the $3'$ region of their target mRNAs by base paring. The seed site, which is located at position 2 to 7 of the miRNA, is usually used for this base paring, and the seed type tends to be one of the most important features. There are also other additional features that can contribute to target recognition and efficacy, such as $3'$ additional paring, site conservation, site accessibility, and cooperability of multiple target sites.

3.6 MicroRNA binding also occurs outside of the 3′ UTR

Although the precise regulatory mechanism of miRISC's binding on the 3′ UTR is unclear, it may cause deadenylation, decapping, and exonucleolytic digestion of the mRNA (59, 60, 61). Moreover, miRISCs bind other regions that reside outside of 3′ UTRs, such as CDS (85), 5′ UTR (85, 86), and promoter regions (10, 11).

Many miRNAs bind CDS regions; however, most of the sites are likely non-functional because ribosomes seem to detach miRISCs from their binding site while moving along the CDS region for translation (3). Nonetheless, there are several evidences that some miRNA target sites in CDS are functional. One example is that rare codons in CDS regions tend to make ribosomes stalled, and miRNA target sites right after these codons can be functional (87). Moreover, some target sites experimentally validated in CDS tend to have one very strong site (88, 89), or multiple sites within optimal distance (90, 91).

Some miRNAs also target 5′ UTR regions, though it is much less common than 3′ UTR and CDS regions (4). One intriguing class of miRNA targets involved in 5′ UTR is miBridge targets (86). The miBridge target is a miRNA target site that has a normal seed site on the 3′ UTR and a 5′ portion paring on the 5′ UTR simultaneously. One possible explanation for the regulation of miBridge is that miBridge is involved in the translational initiation by preventing ribosome scanning through the 5′ UTR (86).

MicroRNA regulation can also occur inside the nucleus despite that miRNA's major regulatory roles are in the cytoplasm (92, 93). The miRNA regulation in the nucleus is likely at the transcriptional level rather than the translational level. Several studies reported that miRNAs cause transcriptional silencing by targeting promoter regions (10, 11).

3.7 Some of coding and non-coding pairs of cis-NATs potentially may have regulatory interactions with miRNAs

A complex locus is a locus that contains several genes that interact among each other (18). Two major classes of complex loci are cis-natural antisense transcripts (cis-NATs)

and bi-directionally promoters (Fig. 3.4) (18). Many complex loci have important regulatory roles, even though the precise mechanism is unknown (94, 95).

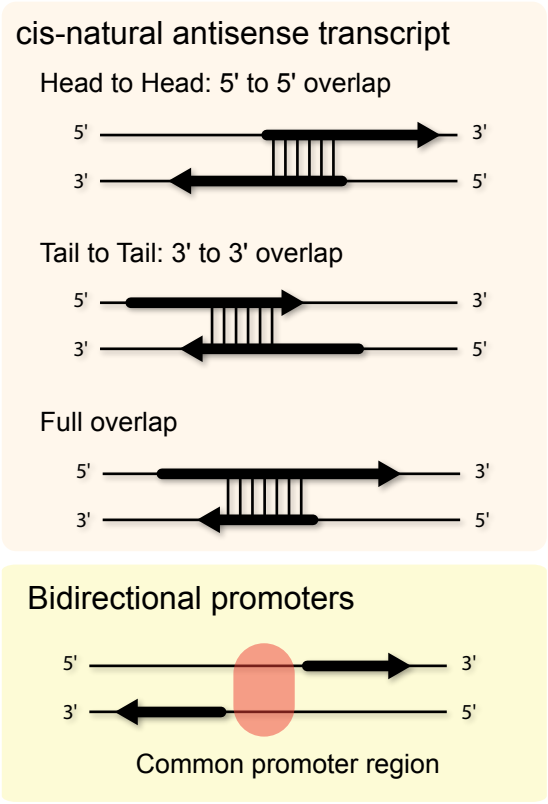

Figure 3.4: Complex loci - Complex loci consist of multiple genes that interact among each other. Two major classes of complex loci are cis-NATs and bi-directional promoters. Cis-NATs can be divided into three categories depending on the directions of the overlaps. "Head to Head" is that sense and antisense transcripts are partially overlapped on their 5′ ends. "Tail to Tail" is that sense and antisense transcripts are partially overlapped on their 3′ ends. "Full overlap" is that one transcript is fully overlapped with the other transcript. Bi-directional promoters reside between two genes arranged head-to-head on opposite strands with less than 1000 base pairs separating their transcription start sites (96)

Cis-NATs is a sense-antisense pair of transcripts that are partially overlapping in the same locus (97). Cis-NATs are relatively common in many species, and they are quite abundant in human (18, 98, 99). Three major classes, or orientations, of cis-NATs are "head to head", "tail to tail", and "full overlap" (Fig. 3.4) (100). Although the molecular mechanism of cis-NAT regulation is poorly understood, three models may

explain the potential cis-NAT regulation (100). The first model is the transcriptional collision model (100), which can be the main mechanism for "head to head" cis-NATs. The second model is the dsRNA formation model of sense and antisense transcripts (100). One study showed that an endogenous ncRNA derived from a cis-NAT pair regulates its anti-sense transcript of a protein coding gene in plants (101). The third model is that cis-NATs are involved in epigenetic regulation. This model is based on the evidence that some ncRNAs are involved in the modification of chromatin structure and DNA methylation in the promoter region (102, 103).

Bi-directional promoters reside between two genes arranged head-to-head on opposite strands with less than 1000 base pairs separating their transcription start sites (96). Many pairs from bi-directional promoters are co-expressed, but some are antiregulated (96). Bi-directional promoters are abundant; for instance, they represent approximately 10% of all the genes in human (96).

The main objective of the third sub-goal: *miRNA and other ncRNAs* is to investigate potential miRNA interactions with other ncRNAs. Of these classes of complex loci, some miRNAs potentially interact with ncRNA:mRNA pairs of cis-NATs. In this case, miRNAs indirectly regulate the expression of the mRNAs in cis-NATs through directly regulating their paired ncRNAs (Fig. 3.5). Bi-directional promoters may also have ncRNA:mRNA pairs that potentially involve miRNAs regulation. Annotated data of bi-directional promoters are available for mRNA:mRNA pairs (18), but there are few reliable data for ncRNA:mRNA pairs of bi-directional promoters. Therefore, we excluded bi-directional promoters and focused on cis-NATs, especially on ncRNA:mRNA pairs of cis-NATs, in the third sub-goal: *miRNA and other ncRNAs*

3.8 Chromatin associated RNAs are potentially associated with the modification of chromatin structure

Chromatin associated RNAs (CARs) are experimentally validated non-coding RNAs that can bind a part of the chromatin directly (104). Since CARs affect their host and neighboring genes (104), CARs can be seen as a class of complex loci. CARs are likely involved in the regulation of chromatin structure by recruiting chromatin-modifying complexes (Fig. 3.6). This scenario is supported by the evidence that long ncRNAs regulate chromatin modification by guiding chromatin remodeling complexes to specific

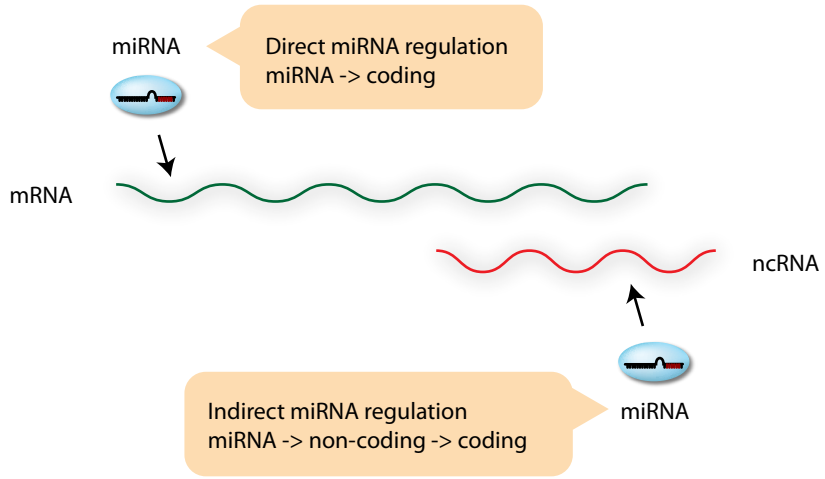

Figure 3.5: miRNA regulation on cis-NAT - The figure illustrates two different potential modes of miRNA regulation on cis-NATs. Direct miRNA regulation is a normal regulatory mode of miRNAs, in which miRNA binds on the 3′ UTR of mRNA. Indirect mode is that miRNA regulates the protein coding mRNA in cis-NAT indirectly through binding the non-coding transcript.

genome loci (105, 106). In addition to long ncRNAs, RNAi is also known to have roles in the regulation of chromatin structure. For instance, in fission yeast, CARs serve as assembly platform to the RNA-induced initiation of transcriptional gene-silencing (RITS) complex. In this case of fission yeast, siRNAs associate with AGO1 and guide the RITS complex to CARs (103).

Similar to ncRNA:mRNA pairs of cis-NATs, CARs potentially have interactions with miRNAs, though there is currently no strong evidence to support this. Therefore, we chose cis-NATs and CARs to investigate their potential interactions with miRNAs in the third sub-goal: *miRNA and other ncRNAs*.

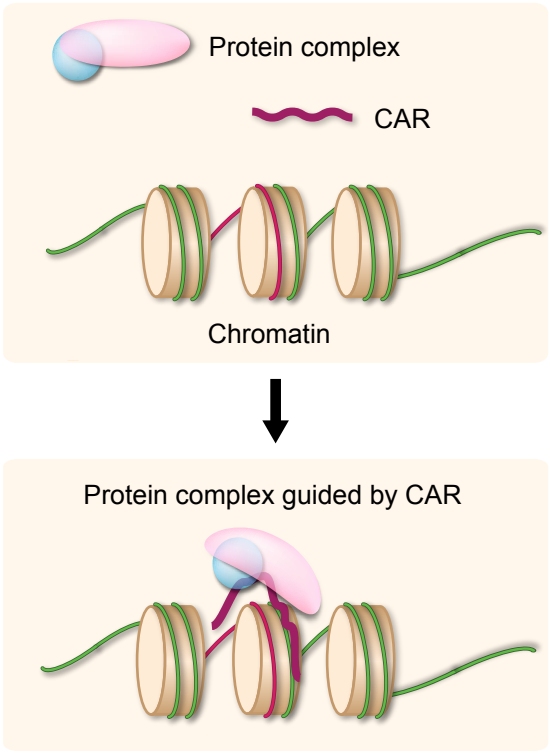

Figure 3.6: CARs - Upper panel shows a protein complex, a CAR, and nucleosomes. The complementary DNA region to the CAR is indicated in red. Lower panel shows one example of the CAR regulation to the chromatin. The CAR acts as a guide for the protein complex that can modify the chromatin structure.

4

High-throughput biological experiments

Emerging high-throughput technologies have enabled genome-wide analyses of various biological data, such as different cell lines, tissues, and species, under different conditions. This chapter explains several high-throughput technologies for transcriptomic and proteomic analyses used in our research. We used both microarray and quantitative proteomics data to achieve all three sub-goals, but we used next generation sequence data only for the second and third sub-goals: *miRNA high-throughput experiments*, and *miRNA and other ncRNAs*.

4.1 One microarray experiment can detect thousands of gene expressions simultaneously

Microarrays can capture the expression levels of many genes in a single set of experiments. The microarray technology enables a high-throughput transcriptome-wide analysis, which is hard to achieve with other techniques for gene expression analyses, such as Northern blot (107) or reverse transcription quantitative polymerase chain reaction (RT-qPCR) (108, 109). The Northern blot is a qualitative but low-throughput technique that requires the use of electrophoresis and large amounts of the input RNA (110). RT-qPCR can achieve higher throughput levels with less amount of the input RNA than Northern blot, but the throughput remains on the order of hundreds of known transcripts at a time (110, 111).

Although several types of microarray technologies exist, the DNA microarray is usually used for the transcriptome analysis. One microarray chip usually consists of thousands of spots, and each spot contains DNA oligonucleotides of a specific sequence (112). These DNA oligonucleotides are suitable for hybridization with DNA or RNA isolated from cells. Isolated RNAs are chemically labeled before hybridization. These labels, such as fluorescence dyes, are used to detect signal intensities of spots and determine the relative mRNA abundance among samples. One of the major DNA microarray applications is to measure the relative difference of mRNA abundance between two samples. For example, Figure 4.1 shows the procedure of a two-color microarray system (112). Another popular microarray design is one-color microarray system, and data quality is essentially equivalent between one- and two-color approaches (113).

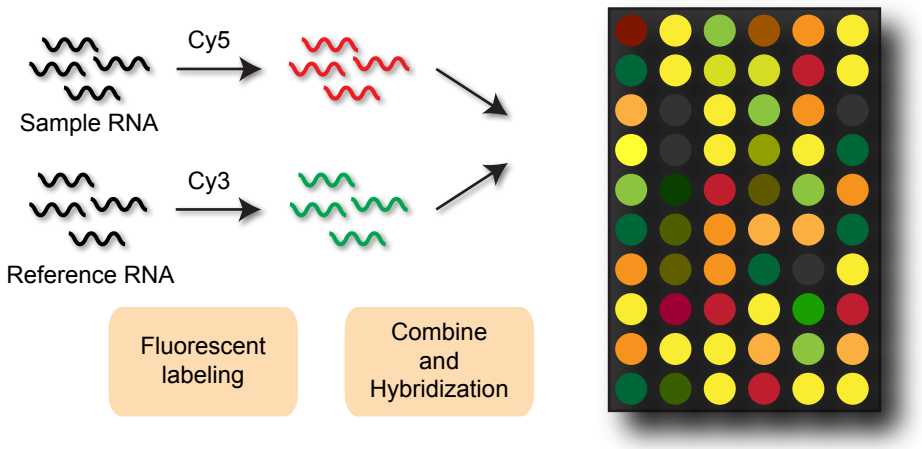

Figure 4.1: Microarray procedure - Two samples are compared by a two-color DNA microarray. Sample RNA represents a sample of interest, whereas Reference RNA represents a control sample for comparison. Sample RNA is dyed red by Cy5, and Reference RNA is dyed green by Cy3. These two samples are combined and hybridized with DNA oligonucleotides on the chip. Color intensity is measured to estimate relative RNA expression levels of the two samples.

After measuring the intensity levels, the raw data usually go through normalization and transformation (114), and this whole process is usually called the pre-processing of microarray data. The pre-processing of the raw data is very important to reduce the noise in each sample at the local level and among multiple samples at the global level. In addition to detecting relative RNA expression levels, the DNA microarray can be

used for other applications, such as the detection of single-nucleotide polymorphisms (SNPs), and alternative splicing events (110, 115).

Two major drawbacks of microarray technologies are the limited ability to detect novel transcripts and noisy data even after pre-processing (110). Nonetheless, the microarray technology is still widely used for many transcriptomics analyses because of the ability to measure the expression of thousands of genes simultaneously at relatively low cost.

4.2 The next generation sequencing methods are faster and more cost-effective than Sanger sequencing

As an alternative approach to the microarray technology, DNA sequencing approaches are also widely used for transcriptome analysis. The advantage of these sequencing methods is the ability to identify new transcripts and measure the abundance of transcripts directly (110). First generation sequencing methods relied on the Sanger method (116). Since the original Sanger method uses the full-length complementary DNA (FLcDNA), it involves a complex *in vivo* cloning step that usually results in very high cost. Therefore, the original Sanger method is normally limited only to novel transcript discovery and annotation (110).

Two examples of Sanger method applications are expressed sequence tag (EST) (117) and Serial Analysis of Gene Expression (SAGE) (118). Both approaches use shorter tags, which are short sub-sequences of the cDNA sequence, rather than the FLcDNAs. ESTs are short tags generated from either $3'$ or $5'$ end of a cDNA clone. Even though the sequencing cost is reduced by ESTs compared with FLcDNAs, it is still too expensive for the whole transcriptome analysis (110). SAGE is a method that uses short tags generated from $3'$ ends of mRNA transcripts. SAGE is suitable for estimating transcript abundance due to high redundancy of sequencing reads (110). However, SAGE is still costly for the transcriptome analysis because it still relies on labor intensive *in vivo* cloning procedures (110).

The next generation, or the second generation, sequencing methods have substantially improved upon the Sanger method. They produce millions of short reads in a relatively short period of time (119) depending on several criteria, such as read length, sequence coverage, and the size of the genome of interest (Table 4.1). The reads are assem-

bled computationally afterwards if necessary. The major contribution to this enhancement is to parallelize the sequencing process (120), though other features, such as the usage of PCR-based amplification instead of costly and labor intensive *in vivo* cloning, also contribute (110). Three popular commercially available next generation sequencing technologies are, Roche 454 (`http://454.com`), Illumina (`http://www.illumina.com`), and Applied Biosystems SOLiD (`http://www.appliedbiosystems.com`) (Table 4.1). The second generation sequencing can be used for all the applications that are based on the Sanger method, including EST and SAGE. Some applications that are based on the second generation sequencing are explained in the next section.

Table 4.1: Next generation sequencing technologies - The table shows three examples of commercially available next generation sequencing technologies with specifications obtained from their corresponding web sites as of March 2011. "bp" and "Gb" represent base pairs and giga base pairs respectively. The product type with the best specification is selected for each technology. Both 2 × n and n × m represent the read length of the pair end approach where n and m are read lengths in base pairs.

	Roche 454	**Illumina**	**SOLiD**
Product type	GS FLX Titanium	HiSeq 2000	5500xl
Read length	400 bp	1 × 35 bp	75 bp
		2 × 50 bp	75 bp × 35 bp
		2 × 100 bp	60 bp × 60 bp
Run time	10 hours	1.5 days (1 × 35)	1 day (75)
		4 days (2 × 50)	7 days (75 × 35)
		8 days (2 × 100)	7 days (60 × 60)
Throughput per day	1 Gb	25 Gb	20-30 Gb

Third generation sequencing techniques will be available in the near future. The main feature of the third generation is the ability to sequence the whole single molecule instead of breaking down the molecule into short reads, therefore the read lengths should be much longer than those of the second generation sequencing technologies. Several strong candidates that may lead the third generation sequencing are Pacific Bioscience SMRT Sequencing (`http://www.pacificbiosciences.com`), Oxford Nanopore

technologies (`http://www.nanoporetech.com/`), and Life technologies Single Molecule Sequencing (`http://www.lifetechnologies.com`).

4.3 The second generation sequencing technologies can cover a wide range of applications

The second generation sequencing can be used in many different applications because of its high-throughput and cost effectiveness. For example, the applications can be transcript rearrangement discovery, single-nucleotide variation profiling, and non-coding RNA discovery (110). Two such applications, RNA-Seq and CLIP (Cross-Linking and ImmunoPrecipitation)-Seq, are very powerful and useful for transcriptomic analyses.

RNA-Seq or the whole transcriptome shotgun sequencing (WTSS) is a technique that uses the second generation sequencing technology to produce sequence reads at the whole transcriptome level (121, 122). Since the second generation sequencing technology can yield sufficient sequencing depth, which represents the total number of sequence reads generated from a sequencing library (110), RNA-Seq can be used for gene expression profiling with high accuracy.

CLIP-Seq (123), also called HITS-CLIP (High throughput sequencing CLIP) (124), is a technique that employs three important steps, cross-linking, immunoprecipitation, and next generation sequencing. It can be used to tag and pull-down RNA-interacting proteins of interest and infer the interactions between RNAs and RNA-binding proteins. Firstly, RNA binding proteins and their target RNA regions are cross-linked by ultraviolet (UV) light, and the antibodies for the proteins are used for immunoprecipitation (125). Subsequently, the RNA transcripts pulled down with the proteins go through the second generation sequencing procedure (123). RIP (Ribonucleoprotein ImmunoPrecipitation)-Seq is similar to CLIP-Seq (126), but it uses chemical cross-linkers such as formaldehyde instead of UV cross-linking (127, 128). This cross-linking is reversible, and it is subject to potential reassociation between RNAs and RNA-binding proteins after cell lysis in some cases (129).

4.4 Liquid chromatography-tandem mass spectrometry is a powerful tool to analyze quantitative proteomics

In recent years, several new technologies for the identification and quantification of proteins have emerged. Most of them are based on mass-spectrometry (MS), which is a technique that ionizes molecules and measures the mass-to-charge ratio by detecting them in an electromagnetic field (130). Although there are many variants of MS-based technologies (131), the Liquid chromatography-tandem mass spectrometry (LC-MS/MS) with stable isotope labeling with amino acids in cell culture (SILAC) (132) approach is widely used in detecting protein expression profiles.

LC-MS/MS uses high-performance liquid chromatography that can separate a mixture of molecules with very small particles and a high pressure before the MS/MS phase. MS/MS, or tandem mass spectrometry, involves two steps of MS selections. The first MS can be used for the quantification of peptides, and the second MS can be used for the identification of the peptides (131).

LC-MS/MS is usually combined with either labeling or labeling-free methods for a quantification approach. For example, SILAC tends to be used for small changes (10%-50%), and isotope tags for relative and absolute quantification (iTRAQ), which is another labeling method, tends to be used for moderate changes (50%-200%) (131). Moreover, a labeling-free method using spectrum counts can be used for large changes (>100%) (131). Among them, SILAC is a simple but very powerful method for quantitative proteomics (133). The SILAC procedure uses two different stable amino acid isotopes, as "light" and "heavy" labels. The relative abundance of proteins can be detected by comparing the intensities of isotope clusters (131, 133).

The coverage of protein identification in the genome is usually less than 10% for higher organisms (134) due to enormous molecular complexity and the dynamic nature of proteins, such as post-translational modifications and protein stability (131). However, the protein coverage of quantification is even lower than the protein identification. One possible explanation for this low coverage is that protein quantification requires much higher data quality, in terms of information content, than protein identification (134).

4.5 Most preprocessed and raw data sets from high-throughput experiments are publicly available

Two major repositories for microarray experiment data are ArrayExpress (`http://www.ebi.ac.uk/arrayexpress`) (135) at European Bioinformatics Institute (EBI), and Gene Expression Omnibus (GEO) (`http://www.ncbi.nlm.nih.gov/geo`) (136) at National Center for Biotechnology Information (NCBI). Both repositories encourage submitters to supply Minimum Information About a Microarray Experiment (MIAME) (137) compliant data. MIAME is a standard for the microarray data formats.

The major repository for the next generation sequencing data is the Sequence Read Archive (SRA), which is operated by the International Nucleotide Sequence Database Collaboration (INSDC) (138). However, due to a rapid growth of next generation sequencing data and budget constrains, NCBI, which is the main member of INSDS, currently accepts limited types and forms of the next generation sequencing data.

There are no central repositories for quantitative proteomics data, but many small and medium scale public repositories are available instead. Some of the examples of such repositories (139) are Proteomics IDEntifications database (PRIDE) (140), the Global Proteome Machine database (GPMDB) (141), and PeptideAtlas (142).

5

Statistical tests and methods

This chapter describes various statistical methods used in our research. We used basic statistical methods, such as parametric tests and correlation, to achieve all three sub-goals of our research, but we used multiple non-parametric tests only for the second and third sub-goals: *miRNA high-throughput experiments*, and *miRNA and other ncRNAs*. Moreover, we mainly used the resampling approach to achieve the third sub-goal: *miRNA and other ncRNAs*.

5.1 Parametric statistics: Parameters and Hypothesis testing

Statistics tests play important roles in biology to analyze different kinds of data from biological experiments. Most analyses in biology use parametric statistics, which can be used only when the data are likely from a known distribution with parameters. The most commonly used parametric distribution is the normal distribution, which has two parameters: μ (mean) and σ^2 (variance). Mean is a measure of central tendency, whereas variance is a measure of spread. Standard deviation (σ), which is the square root of variance, is also a measure of spread. The normal distribution is defined by its probability density function (143) as:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{5.1}$$

The parametric statistics offers various analysis methods, but the most common method is hypothesis testing. Hypothesis testing is that the null hypothesis, denoted

by H_0, is tested to infer whether the alternative hypothesis, denoted by H_1, is true or false. The alternative hypothesis contradicts the null hypothesis in some sense (143), therefore, if H_0 is rejected, H_1 is inferred as "True", whereas if H_0 is accepted, H_1 is inferred as "False".

The p-value is the probability of incorrectly rejecting the null hypothesis when it is true. For example, the p-value 0.05 means that there is 5% chance of rejecting the null hypothesis when it is true. Two significance levels, 0.05 and 0.01, are commonly used as statistically "significant" or "highly significant". Moreover, two types of errors may occur when the null hypothesis is either accepted or rejected (Table 5.1). Type I error is the error of rejecting the null hypothesis when it is true, whereas Type II error is the error of accepting the null hypothesis when it is false. Type I error is more important for hypothesis testing because the p-value is equivalent to the probability of Type I error.

Table 5.1: Four possible outcomes of hypothesis testing - The table shows the four possible outcomes of hypothesis testing with two error types.

	H_0 is true	H_1 is true
Accept H_0	True Negative	False Negative (Type II error)
Reject H_0	False Positive (Type I error)	True Positive

For analysis of biological data, one of the most common methods for hypothesis testing is two sample inference. For example, when two samples, x_1 and x_2, are normally distributed with equal variance, the test statistic t (143) is:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{5.2}$$

where $S = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2)}{n_1+n_2-2}}$, n_1 and n_2 are sample size of x_1 and x_2, and s_1 and s_2 are standard deviation of x_1 and x_2. The test statistics enables to determine the sampling distribution under the null hypothesis, hence the p-value can be calculated from the test statistics. The calculation method of the test statistics varies depending

on the type of distributions and the properties of the samples. In the example above, the test statistic t follows Student's t distribution. This test is called two sample Student's t-test, and it is used when the variances need to be calculated directly from the samples.

5.2 Non-parametric statistical methods: Wilcoxon rank-sum and Kolmogorov-Smirnov tests

The parametric statistical methods are valid only when the samples of interest follow known distributions with parameters. However, the original distributions of samples are quite often unknown, therefore, non-parametric statistical methods should be used in these cases. Non-parametric methods tend to be more robust, and their applicability is much wider than corresponding parametric methods because they need fewer assumptions. However, they require a larger sample size to draw the same conclusion of their corresponding parametric methods because they usually have less statistical power.

One of the most commonly used non-parametric statistical methods is the Wilcoxon rank-sum test (144, 145), which is a two sample non-parametric test when the samples are independent. It uses a ranking procedure, in which individual values are ordered and ranked. There are two approaches, the Mann-Whitney U-test and the normal approximation, to calculate the test statistics for the Wilcoxon rank-sum test. (i) The Mann-Whitney U-test (144) is used to test whether two samples are drawn from the same distribution. The U value for the U-test is calculated from the sum of the ranks and the sample size. For example, the U value for sample x, denoted as U_x, is calculated as $U_x = R_x - n_x(n_x+1)/2$ where R_x is the sum of the ranks of x, and n_x is the sample size of x. (ii) The normal approximation can be used instead of the U-test when the sample size is large enough (>10) for both samples (143). The test statistics T for two independent samples, x and y, is:

$$T = \frac{\left[\left|R_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}\right]}{\sqrt{\left(\frac{n_1 n_2}{12}\right)(n_1+n_2+1))}}, \tag{5.3}$$

where R_x is the sum of the ranks of x, and n_x and n_y are the sample size of x and y.

The Kolmogorov-Smirnov test (K-S test) is a non-parametric statistical method that does not use ranking procedures. For instance, the two sample K-S test is used to infer whether two continuous distributions differ. The K-S test requires two continuous distribution functions, $F(x)$ and $G(y)$ where the two distributions are defined as $X_1 \dots X_m$ with the size m, and $Y_1 \dots Y_n$ with the size n. In this case, however, both distributions are unknown. Therefore, empirical distribution functions, $\hat{F}(x)$ and $\hat{G}(y)$, are used instead. An empirical distribution function is a step function defined as (146):

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \tag{5.4}$$

where $I(X_i \leq x)$ is the indicator function, and is equal to 1 if $X_i \leq x$ and 0 otherwise. The test statistics D for the K-S test (147) is:

$$D = \max_x |\hat{F}(x) - \hat{G}(x)|, \tag{5.5}$$

for the hypothesis of this test:

$$\begin{array}{lll} H_0: & F(x) = G(x) & \text{for all } x, \\ H_1: & F(x) \neq G(x) & \text{for some } x. \end{array}$$

H_0 is rejected at level α when D is too large as in:

$$\frac{mn}{m+n} D > K_\alpha, \tag{5.6}$$

where the critical value of the Kolmogorov distribution, K_α, is found from $P(K \leq K_\alpha) = 1 - \alpha$ (148).

5.3 Resampling: Bootstrap and Permutation test

In statistics, resampling methods treat an observed sample as a finite population (149) and reuse the data of the observed sample. Resampling approaches have gained popularity in recent years because sufficient computational power has become available to make enough random samples to achieve robust statistical analysis (150). Three major applications of resampling are (i) the bootstrapping method as estimating the characteristics of the sample, (ii) the permutation test as exchanging labels to perform significant tests, and (iii) the cross validation approach as validating models by using random subsets. This section briefly explains two such applications, bootstrapping and

permutation tests. Cross-validation is explained in the next chapter as an evaluation method for machine learning.

Bootstrapping (151) is a resampling method that generates random samples from an observed sample with replacement. Sampling with replacement means that a randomly drawn observation should put back in the original sample before drawing the next one (150). Bootstrap is mainly used for estimating population characteristics by collecting the statistics from many resamples.

Permutation tests are non-parametric procedures based on resampling. The tests randomly rearrange the data without replacement to create the sampling distribution of the test statistics under the null hypothesis (150). To illustrate the basic idea of a permutation test, suppose we have two samples x with size m and y with size n. We first pool all the data points from x and y, and randomly draw a point from this pooled set without replacement to make resample controls with size m and n. We then iterate this resampling to make a permutation distribution. The number of resamples depends on a required statistical power, but 1000 is widely used. The p-value is calculated by comparing the parameter of the original observation with the permutation distribution of the parameter (150). For instance, if 14 cases of 999 resamples are larger than the parameter of the original sample, the p-value of one-sided test can be calculated as:

$$\frac{14+1}{999+1} = \frac{15}{1000} = 0.015. \tag{5.7}$$

Adding one to both numerator and denominator of Eq. (5.7) improves the estimate of the p-value. Moreover, Fisher's exact test (152) is a special case of permutation test that is used in the analysis of categorical data, especially for contingency tables with small sample size. For instance, when the Fisher's exact test is used for a 2×2 table, it calculates the exact probability by considering all possible values under the assumption that the margins of the table are fixed (150).

5.4 Multiple comparison tests: Analysis of variance, Bonferroni correction, and False discovery rate

In addition to one and two sample inferences, multisample inference is also important in many biological analyses. Two major approaches for multisample inference are the analysis of variance (ANOVA) and multiple comparison tests.

The ANOVA test concerns the means of several groups, and its hypothesis is:

$$H_0 : \text{all means are equal,}$$
$$H_1 : \text{not all means are equal.}$$

The F test can be used when each group follows a normal distribution, and the test statistics F is:

$$F = \frac{\text{Between Mean Square}}{\text{Within Mean Square}}. \tag{5.8}$$

"Between Mean Square" measures the mean among the groups, whereas "Within Mean Square" measures the mean among individuals within the same group (150).

As for the non-parametric approach, the Kruskal-Wallis test (153) can be used if some group has no specific distribution. It is a non-parametric ANOVA test, and it uses ranking procedures as calculating the sums of the ranks for the groups (150).

Multiple comparisons procedures enable to detect the groups that differ from the others. The most common approach of multiple comparisons is to simply compare all possible pairs by two sample inference, followed by p-value adjustment. The p-value adjustment is critical for multiple comparisons because some differences likely occur just by chance if there is a large number of groups, and every pair of groups should be compared (150). Many p-value correction methods have been developed for various cases, and most of them either change the significance level of the test, α, or consider the false discovery rate (FDR), which is (False Positive) / (False Positive + True Negative).

The Bonferroni correction computes an alternative significance level, α^*, defined as (143):

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}, \tag{5.9}$$

where k is the number of groups. For example, there are 45 possible pairs when $k = 10$, therefore $\alpha^* = \alpha/45$. One critical problem of the Bonferroni correction is to control the overall experimental-wise type I error rate, hence, no significant pairs may be found when k is very large.

The FDR control, or the Benjamini and Hochberg correction (154), is to modify p-values without controlling the overall experimental-wise type I error rate. The FDR aims to control the proportion of false-positive results (143), therefore, several significant pairs will be expected to be found.

5.5 Correlation: Pearson's and Spearman's correlation coefficients

In statistics, the correlation indicates the statistical relationships between two or more samples. The correlation coefficient, which ranges from -1 to 1, represents the degree of correlation. Samples are positively correlated, negatively correlated, and uncorrelated when the coefficient is greater than 0, less than 0, and exactly 0, respectively (143). It is also important to test the significance of correlation by determining whether an observed correlation coefficient is significantly different from zero or not.

Pearson's correlation coefficient, usually denoted as r, indicates the linear relationships between two samples that follow normal distribution. For example, two samples, X and Y, have individual observations represented as x_i and y_i where i = 1, 2, ..., n. The Pearson's correlation coefficient, r_{xy}, is (143):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{5.10}$$

This is equivalent to:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\text{sample covariance between } x \text{ and } y}{(\text{sample standard deviation of } x)(\text{sample standard deviation of } y)}.$$

Spearman's correlation coefficient, denoted as ρ, is a non-parametric method. Hence, it can be used when the distributions are unknown. The calculation of ρ is similar to that of r, but the ranks are used instead of the actual observation values (143).

5.6 Regression analysis: Multivariate linear regression

Regression analysis is an important statistical method with biological data because it identifies the characteristics and relationships among multiple factors (155). Many types of regression analysis exist depending on different criteria such as univariate versus multivariate, or linear versus non-linear, for instance.

Multivariate linear regression can be performed to study the effect of multiple variables and their linear relationships in the data. The linear regression model relating y to x_1, ..., x_k, is (143):

$$y = \alpha + \sum_{j=1}^{k} \beta_j x_j + e, \tag{5.11}$$

where e is an error term that is normally distributed with mean 0 and variance σ^2. The main goal of the regression analysis is to minimize e and estimate the best α and β to fit this model.

The goodness of fit for a regression model indicates how well the observed data fit the predicted model. One of the approaches to measure the goodness of fit for multiple regression models is to perform residual analysis (143). Moreover, many procedures of regression analysis overlap with those of machine learning. Therefore, several machine learning evaluation methods are also useful to evaluate regression models. Some of these evaluation methods for machine learning are explained in the next chapter.

6

Machine learning theory and Support vector machine

The support vector machine (SVM) is a machine learning technique that has been applied in numerous bioinformatics domains successfully in recent years (156). We used SVM as a machine learning model to achieve the first sub-goal of our research: *miRNA target prediction*. We built a binary classification model with both linear and non-linear approaches. Binary classification predicts only two class labels, positive/true or negative/false. This chapter describes the theoretical background for SVM, and data preparation and evaluation methods mainly for binary classification, followed by other machine learning algorithms for comparison.

6.1 Machine learning: Supervised and Unsupervised

Machine learning (ML) techniques have two major paradigms, supervised and unsupervised learning. Supervised learning is used for discriminant analysis and regression analysis (157), and it requires a training process. The training data consists of multiple feature vectors and the class labels. The main purpose of the training process is to make a classifier that can predict appropriate class labels from the feature vectors of the test data. The test data consists of the same feature vectors as in the training data, but it has no class labels. Unsupervised learning requires no training process, and it categorizes unlabeled data. The main application of unsupervised learning is clustering. The aim of clustering is to divide the data into groups (clusters) using some

measures of similarity (157).

6.2 SVM: Theory

SVM is a state-of-the-art supervised machine learning method introduced by Boser, Guyon, and Vapnik in 1992 (158). SVM is a linear binary classification method based on the Structural Risk Minimization (SRM) principle (159). Two essential ideas of SRM are the bound on the generalization performance of a learning machine and the Vapnik Chervonenkis (VC) dimension (160).

The aim of SRM is to find a hypothesis h that has the guaranteed lowest probability of error $Err(h)$ from a hypothesis space H (161). In other words, SRM finds the best machine learning model, α, that has lowest test error rate, $R(\alpha)$, where $R(\alpha)$ is equivalent with $Err(h)$. The upper bound of the test error can guarantee the performance of a learning machine, and the bound holds with a probability of at least $1 - \eta$ for a given training sample with n examples (162):

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{d(\log(2n/d) + 1) - \log(\eta)}{4}}, \tag{6.1}$$

where $R_{emp}(\alpha)$ is a training error rate, and d is a VC dimension.

The VC dimension is a measure of capacity for the data point separation by hyperplanes, and this separation is called shattering. A hyperplane in Euclidean space can separate the space into two half spaces, and it is a subset of $n - 1$ dimension for an n-dimensional space. For example, a straight line is a hyperplane of a two-dimensional Euclidean space, whereas a flat-plane is a hyperplane of a three-dimensional Euclidean space. The VC dimension varies depending on the selection of a machine learning model. For example, the VC dimension of the set of oriented hyperplanes in $\mathbf{R}^n$ is $n + 1$ for a simple linear binary classifier, such as perceptron. Accordingly, Eq. 6.1 reflects a trade-off between the training error, $R_{emp}(\alpha)$, and the complexity of hypothesis space estimated by the VC-dimension of a learning machine (161).

SVMs solve this trade-off problem by keeping the VC-dimension low through maximizing the margin of boundaries. A SVM can be defined as a linear binary classifier:

$$f(\mathbf{x}) = sign(\mathbf{w^T x} + b) = \begin{cases} +1 & \text{if} \quad \mathbf{w^T x} + b > 0 \\ -1 & \text{else} \end{cases}, \tag{6.2}$$

where w is a weight vector and b is a threshold. The margin of this classifier, δ, is a length between a boundary hyperplanes, either $\mathbf{w^T x} + b = +1$ or $\mathbf{w^T x} + b = -1$, and the optimal hyperplane, $\mathbf{w^T x} + b = 0$. The margin is calculated as:

$$\delta = \frac{1}{||\mathbf{w}||}. \tag{6.3}$$

Vapnik proved that the VC dimension d for such a classifier defined in Eq. (6.2) is bounded by (159):

$$d \leq \min\left(\left[\frac{\mathbf{R}^2}{\delta^2}\right], N\right) + 1, \tag{6.4}$$

when this classifier is in an N dimensional space, and all example vectors, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, are inside a ball of diameter $\mathbf{R}$. It indicates that a SVM classifier keeps the VC-dimension lower by maximizing the margin of the boundaries between two hyperplanes. This is the mathematical background to guarantee that SVM has an upper bound of the test error rate even with very large feature vectors since the number of feature vectors has very small influence on the VC-dimension with SVM.

6.3 SVM: Linear SVM

The aim of SRM is to find an optimal hyperplane than can maximize the right term of Eq. (6.1). SVM is based on SRM, and its solution is to make a classifier as Eq. (6.2) with the maximum margin of Eq. (6.3). In other words, from many boundaries that can separate two classes (Fig. 6.1A), the SVM finds the optimal hyperplane where the margin of the boundary between two hyperplanes, $\mathbf{w^T x} + b = +1$ and $\mathbf{w^T x} + b = -1$ becomes the maximum (Fig. 6.1B). Because it is difficult to solve this maximum margin problem directly, this problem is usually translated into the quadratic optimization problem (161). Quadratic programming (QP) is a class of optimization algorithms to either minimize or maximize a quadratic function subject to linear constrains. For SVM, finding a solution in the prime form of QP defined in Eq. (6.5) is equivalent to finding the optimal hyperplane with the maximum margin. The training data of this SVM are $(\mathbf{x}_i, y_i)$ for $\forall i$ where $\mathbf{x}_i$ is a feature vector with N dimensions as $\mathbf{x}_i \in \mathbf{R}^N$, and y_i is a class label as $y_i \in \{-1, +1\}$.

$$\begin{array}{ll} \underset{\mathbf{w},b}{\text{minimize}} & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to :} & y_i(\mathbf{w^T x}_i + b) \geq 1, \forall i. \end{array} \tag{6.5}$$

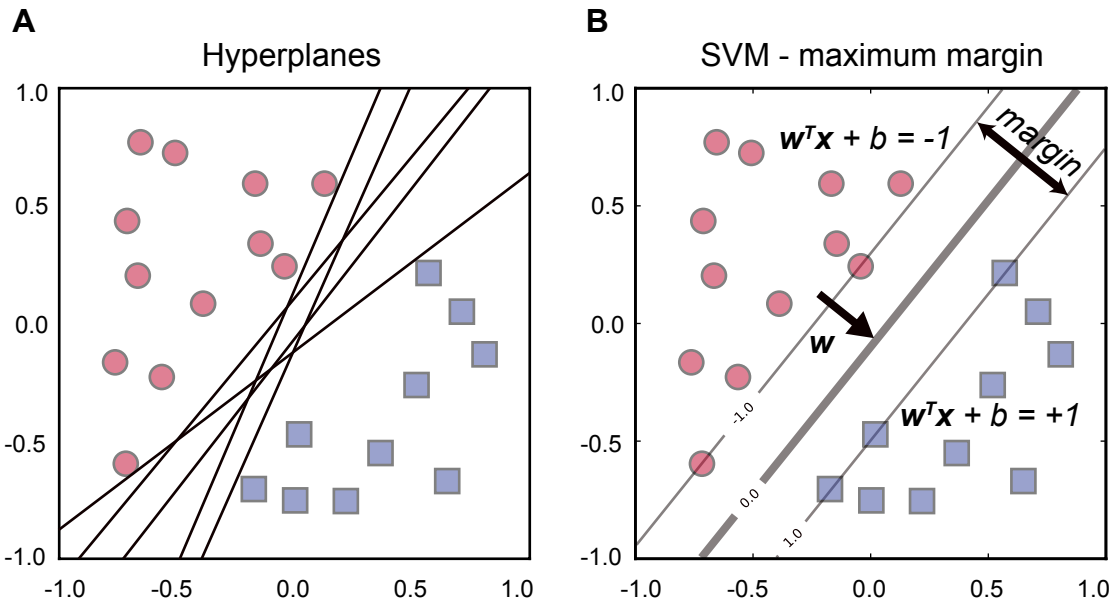

Figure 6.1: SVM hyperplanes and maximum margin - The figure shows the relationship between hyperplanes and the maximum margin of SVM. Red circle and blue square dots represent data points with negative and positive labels. (A) Many hyperplanes can separate two classes. (B) Three lines represent hyperplanes with the optimal hyperplane in the middle. The margin is a distance between two hyperplanes, $\mathbf{w^T x} + b = +1$ and $\mathbf{w^T x} + b = -1$. $\mathbf{w}$ is a weight vector.

However, this prime form of QP (Eq. 6.5) is still numerically difficult to handle (161), therefore, it can be further transformed into the dual form:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \mathbf{x_i^T} \mathbf{x}_j \\ \text{subject to :} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leq \alpha_i \end{aligned} \tag{6.6}$$

Support vectors are vectors such that their corresponding α values are non-zero in this dual form. SVMs only use these support vectors when classifying the test data. The dot product of $\mathbf{x_i^T} \mathbf{x}_j$ can be used for the kernel trick that enables SVMs to solve non-linear problems (163).

In many cases, SVMs can find no hyperplanes that separate two classes, therefore, the slack variables, ξ, are introduced to allow some misclassified data points (161). The SVM with ξ_i is called a soft-margin SVM (164), and its prime form is:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i \\ \text{subject to :} \quad & y_i(\mathbf{w^T} \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i, \end{aligned} \tag{6.7}$$

where C is the cost factor that controls the training error rate. Small C allows many misclassified points (Fig. 6.2A), whereas large C allows few misclassified points between the boundaries (Fig. 6.2B).

6.4 SVM: Non-linear SVM

SVMs can use a kernel function to solve non-linear problems. When the feature vectors are mapped to a high dimensional ($>d$) feature space, $\mathcal{H}$, from the original d-dimensional feature space, $\mathbf{R}^d$, a non-linear separation in $\mathbf{R}^d$ becomes a linear separation in $\mathcal{H}$ (Fig. 6.3) (165). The non-linear mapping function is defined as:

$$\Phi : \mathbf{R}^d \mapsto \mathcal{H}. \tag{6.8}$$

The dot product, $\mathbf{x_i^T} \mathbf{x}_j$, in Eq. (6.6) can be replaced by a kernel function defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x_i^T}) \cdot \Phi(\mathbf{x}_j). \tag{6.9}$$

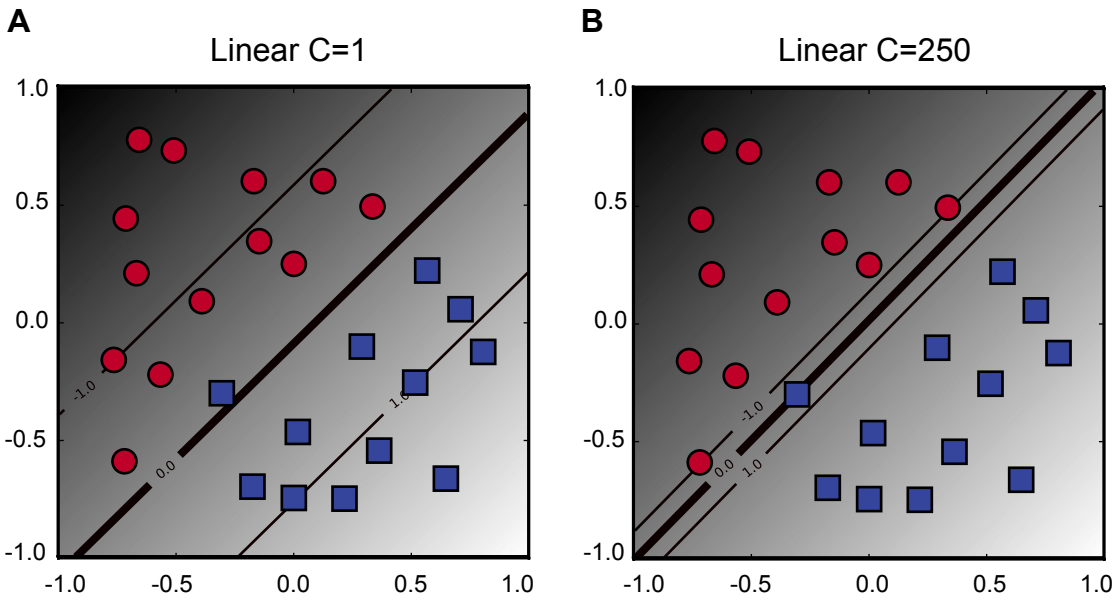

Figure 6.2: Linear kernel with soft margin - The figure shows two examples of different soft margin constants. Red circle and blue square dots represent data points with negative and positive labels, respectively. Three lines represent hyperplanes with the optimal hyperplane in the middle. The gray scale in the background indicates discriminant values. The darker indicates the smaller in negative values, whereas the lighter indicates the greater in positive values. (A) $C = 1$. (B) $C = 250$. In this example, SVM(A), the SVM in panel A, misclassifies one point, whereas SVM(B), the SVM in panel B, classifies all points correctly. However, SVM(A) found a better hyperplane between the overall trends in circle and square classes than SVM(B). SVM(B) has a much smaller margin to the circle class than SVM(A).

Two commonly used non-linear kernel functions (157) are Gaussian Radial Basis Function (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2), \tag{6.10}$$

and Polynomial:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\mathbf{T}} \cdot \mathbf{x}_j + 1)^d. \tag{6.11}$$

These functions are applicable as SVM kernel functions because they can be cast in terms of dot products in Eq. (6.9) (166). Gaussian RBF has a parameter, gamma γ, and Polynomial has a parameter, degree d. These parameters are called kernel parameters, and their optimal values are usually unknown. A common practice to find optimal values for the kernel parameters is to use k-fold cross validation.

Figure 6.3 shows four examples of both Gaussian RBF and Polynomial in two-dimensional space. The gamma parameter in Gaussian RBF represents an RBF width, which is sometimes referred to as $1/2\sigma^2$, and a larger value means a smaller radius. For example, Figure 6.3A has gamma 0.5, and it is less specific and has a larger radius than Figure 6.3B with gamma 4.0. The degree parameter d in Polynomial is a degree in a polynomial function, as in $f(x) = a_d x^d + a_{d-1} x^{d-1} + \cdots + a_2 x^2 + a_1 x + a_0$. Figure 6.3C and D show Polynomial kernels with degree 2 and 3, respectively.

6.5 Classifier evaluation: Confusion matrix and Receiver operating characteristics

A binary classification is a type of classifications that predicts two class labels. To evaluate the performance of a binary classification model, one approach is to use the performance measures derived from the 2×2 confusion matrix that shows four possible outcomes with actual and predicted values (Table 6.1) (167). Another approach is to use Receiver operating characteristics (ROC) (168). The confusion matrix requires only outcome labels as True or False, whereas ROC requires both outcome labels and discriminant values.

Standard performance measures that are derived from the confusion matrix are Accuracy (ACC), Error rate (ERR), Sensitivity (SN), Specificity (SP), Positive predictive value (PPV), and Negative predictive value (NPV), and their equations are summarized in Table 6.2. SN, SP, and PPV are also equivalent to True positive rate (TPR), True negative rate (TNR), and Precision (PRC), respectively.

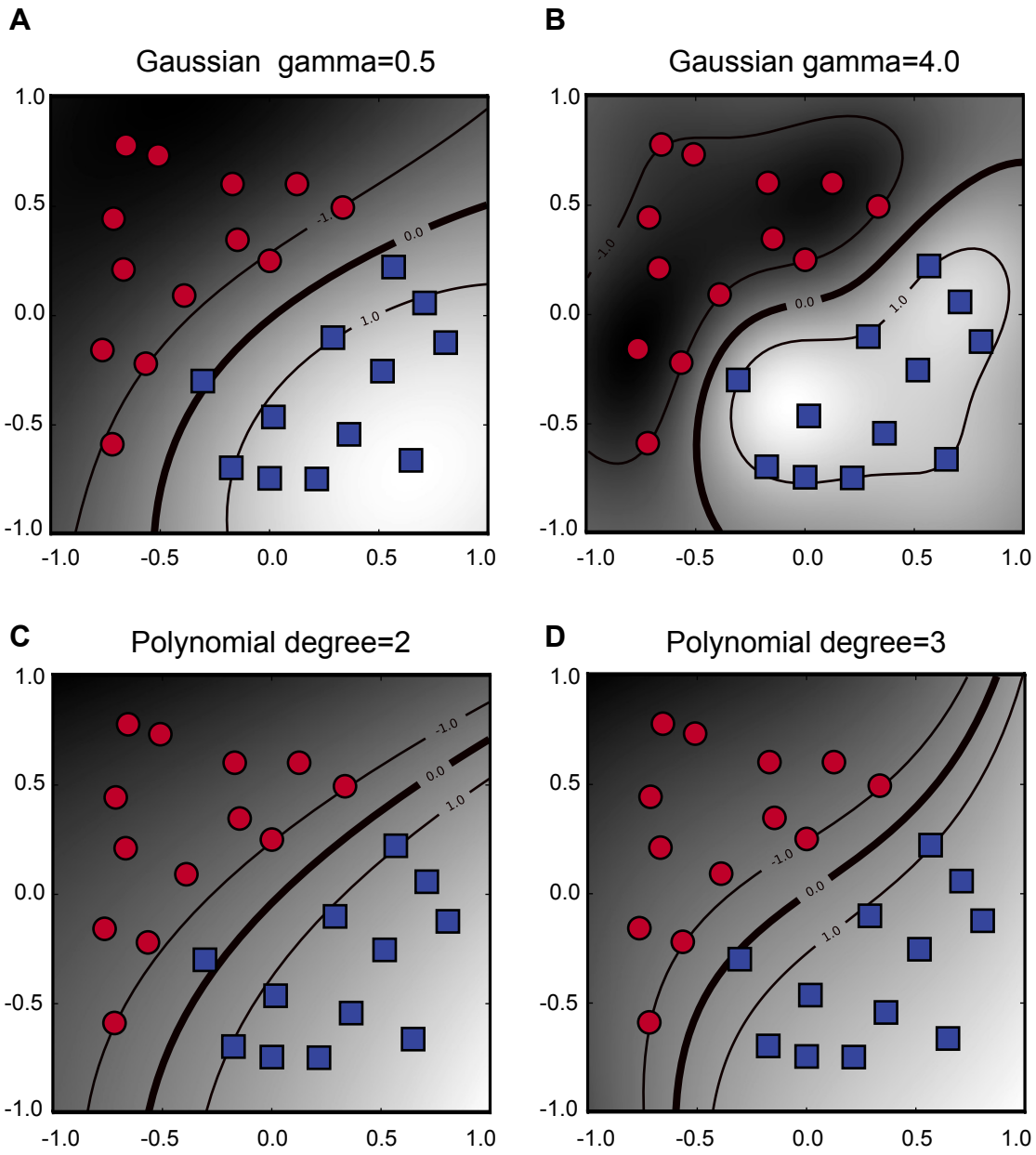

Figure 6.3: Non-linear kernels - The figure shows four examples of two different non-linear kernel functions in two-dimensional space. Each kernel has two plots with different parameter values. Red circle and blue square dots represent data points with negative and positive labels, respectively. Three lines represent hyperplanes with the optimal hyperplane in the middle. The gray scale in the background indicates discriminant values. The darker indicates the smaller in negative values, whereas the lighter indicates the greater in positive values. The cost factor C is 10 for all four kernels. (A) Gaussian RBF with $\gamma = 0.5$. (B) Gaussian RBF with $\gamma = 4.0$. (C) Polynomial with $d = 2$. (D) Polynomial with $d = 3$.

Table 6.1: Confusion matrix - The table shows four possible outcomes of binary classification.

		Actual value	
		Positive (P)	Negative (N)
Prediction outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 6.2: Performance measures from confusion matrix - The table shows the term and equation of six performance measures from the confusion matrix.

Performance measure	Equation
Accuracy (ACC)	(TP + TN) / (P + N)
Error rate (ERR)	(FP + FN) / (P + N)
Sensitivity (SN)	TP / P
Specificity (SP)	TN / N
Positive predictive value (PPV)	TP / (TP + FP)
Negative predictive value (NPV)	TN / (TN + FN)

A ROC graph is a plot that shows False Positive Rate (FPR) or 1 - SP on the x axis and TPR on the y axis (Fig. 6.4). Many classifiers produce scores or discriminant values that can be used to adjust TPR and FPR. A ROC graph uses these adjusted TPR and FRP to draw curves by changing the threshold values of the scores. A single ROC point is drawn if a classifier has no such adjustment mechanism. A ROC graph contains all information in the confusion metrics, since FN is the complement of TP, and TN is the complement of FP (168). The area under the ROC curve (AUC) is a performance measure to evaluate the ROC curves. The AUC indicates a perfect prediction and a random prediction when it is 1.0 and 0.5, respectively (Fig. 6.4A). In many cases, the ROC with AUC is a more adequate method to evaluate the classifier performance than single-number measures, such as Accuracy (ACC) (169, 170).

It is also important to evaluate the trend of ROC curves. For instance, Figure 6.4B shows two classifiers, Classifier1 and Classifier2, that yield the same AUC scores. Classifier1 is a better classifier despite the same AUC values, because it has a better early retrieval performance, which means it has higher TPR when its SP is also high. The area marked orange in Figure 6.4B is the important region to estimate the early retrieval performance. Two variants of ROC, ROC_{50} (171) and concentrated ROC (CROC) (172) are especially useful when measuring the early retrieval performance. They are similar to evaluating the ROC curves in the high SP area as in Figure 6.4B, however, the AUC values of ROC_{50} and CROC can directly indicate the early retrieval performance.

6.6 Training and Test data: Single dataset hold-out and k-fold cross validation

Supervised learning requires a test dataset for performance evaluation. A common problem in machine learning is overfitting, where the model overfits the training data, and it is generalized poorly to unseen data. Therefore, elaborate test procedures are important to maximize the performance and avoid overfitting simultaneously. Two major approaches to separate the test data set from the training data set are single dataset hold-out and k-fold cross validation.

The single dataset hold-out testing is the simplest and most intuitive way to make a test dataset. The sample S_n is divided randomly into two parts, S_l for training and

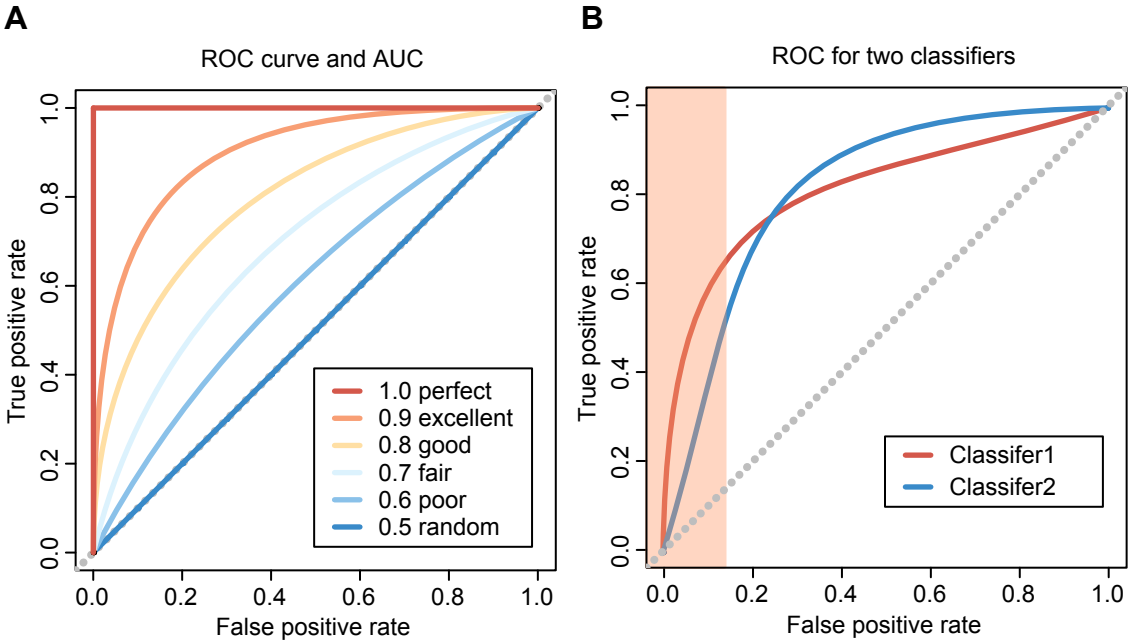

Figure 6.4: ROC curves and AUC scores - The figures show six ROC curves with corresponding AUC scores and an example ROC plot with two classifiers. (A) The best AUC score is 1, and the worst score is 0. The random guess would yield the AUC score 0.5, and the predictions are opposite to expected when the AUC score is between 0 and 0.5. Six ROC curves have different AUC scores between 0.5 and 1.0 to show different prediction performances. (B) Two classifiers, Classifier1 and Classifier2 have the same AUC scores. The orange area in the left part of the plot indicates the critical area on evaluating ROC curves in terms of the early retrieval performance.

S_k for testing, where the sample size n is equal to $l + k$ (161). S_k is treated as an independent testset, and it should never be used for training. Selecting values of l and k is a trade off, because larger l results in smaller bias of the classifier, and smaller k increases the variance for the evaluation (161).

Cross validation is the most popular method because the whole dataset is trained and tested. There are several versions of the cross validation approach. Leave-one-out is a good performance estimator in many cases, but it demands very high computational power and it is also sensitive to the data structure. With leave one-out, one data point is removed from the training dataset for testing, and the test procedure continues until each data point is tested. To reduce the computational time, k-fold cross validation is commonly used instead of the leave-one-out approach. In k-fold cross validation, the training set is partitioned into k folds. One fold from the k folds is used as the test dataset, where the remaining dataset with k-1 folds is used for training (161). 10-fold cross validation with $k = 10$ is widely used for the performance evaluations of classifiers.

Analyzing unbalanced datasets is very common in bioinformatics. For example, some datasets are dominated by negative records with very few positive records, therefore the positive:negative ratio is unbalanced. Unbalanced datasets can present a challenge for any machine learning algorithm to achieve good performance (173, 174). SVM can deal with unbalanced data by assigning different soft-margin constants to each class label (175). However, it is important to consider this imbalance when making training and test datasets. One approach to solve this imbalance is to use stratification. For example, a stratified k-fold cross validation selects k-fold datasets by sampling data from each subpopulation (stratum), which is a subset of the data with the same class label, independently.

6.7 SVM: Data pre-processing

It is important to process data before training and testing. Two common approaches that may improve the SVM performance are categorical data conversion and scaling.

SVM requires numerical feature vectors, therefore, categorical data need to be transformed. SVM usually achieves better performance when using m feature vector components to represent an m-category feature instead of a single component (176).

For example, a feature for a single RNA nucleotide {A,C,G,U} can be represented as (0,0,0,1), (0,0,1,0), (0,1,0,0) and (1,0,0,0).

Scaling is very important for various machine learning algorithms including SVM (177). Two common ways of scaling are linear scaling (176) and standardizing (175). With the linear scaling, all vector components are linearly transferred into the same range, for example, [-1, +1] or [0, 1], whereas with the standardizing, they are normalized by their mean and standard deviation.

6.8 SVM: Model selection

The optimization of a classifier is an important phase to maximize the classifier performance. A common practice of model selection with SVM is to evaluate the classifier performance with different kernels and their corresponding parameters. The linear kernel has only one parameter, the soft-margin C, but non-linear kernels tend to have more parameters. The standard method of parameter optimization with two parameters is via grid-search (178). For example, the Gaussian kernel has two parameters (C, γ), and the grid-search is performed by gradually changing the values of both parameters.

6.9 SVM: Multiclass and Regression

In some cases, classifications involve more than two classes. Although some machine learning algorithms are strictly limited to binary classification, there are several SVM approaches that can handle multi-class problems (179). One example of such SVM multi-class approaches is a one-against-the-rest strategy (161). This strategy decomposes a multi-class problem into multiple independent binary classifications (179).

A version of SVM for regression is called Support Vector Regression (SVR) (163). The major difference between SVM and SVR is that the labels are real numbers for SVR rather than categorical data. The optimization problem of SVR is very similar to that of SVM with the prime form as:

$$\begin{array}{ll} \underset{\mathbf{w},b}{\text{minimize}} & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to :} & \begin{cases} \mathbf{w^T}\mathbf{x}_i + b - y_i \leq \varepsilon \\ y_i - \mathbf{w^T}\mathbf{x}_i - b \leq \varepsilon \end{cases}, \end{array} \tag{6.12}$$

where ε is used to control the errors. SVR can also be transformed to the dual form and handle soft-margin and kernel functions (163).

6.10 Other supervised learning algorithms: Decision tree, Artificial neural network, Naive Bayesian, and k-nearest neighbor

Some popular and widely used supervised learning algorithms are Decision tree (180), Artificial Neural Network (ANN) (181), Naive Bayes (NB) (182), and k-nearest neighbor (k-NN) (183). These four algorithms are often used for comparisons with SVM, and they are also introduced as supervised learning algorithms together with SVM (157, 161, 165, 184, 185).

Decision tree learning uses a tree-like hierarchical graph with multiple nodes. Three different types of nodes are the root as the first node, internal nodes, and the leaves as terminal nodes. A predicted class label can be obtained when the decision making process reaches a leaf, because each leaf contains a probability distribution over the class labels as the results of training (186). Some learning algorithms, such as Random Forests (187), can combine multiple decision trees. The Random Forests algorithm makes many different decision trees randomly, and it predicts the class labels by combing the outputs of these decision trees (186).

ANN is a machine learning method that is inspired by the structure and functionalities of biological neural networks in the brain (165, 188). The central nervous system (CNS) in the brain has interconnected networks of neurons that communicate among each other by sending electric pulses through axons, synapses and dendrites. ANN consists of nodes, or "neurons", that are connected together to form a network that mimics the network in the CNS. The single layered perceptron, which is the simplest form of ANN, is a simple feedforward network (189). The single layered perceptron can only solve linearly separable problems (188). The multi-layered perceptron has been developed to solve non-linear problems (181), and it has an input layer, an output layer, and one or more hidden layers in-between. The multi-layered perceptron with the error back-propagation method (190), which is an algorithm that aims at optimizing the weight values by minimizing the errors from the output layer to the hidden layers, enables fairly complex neural networks to solve non-linear problem (181).

The NB classifier is a type of statistical learning algorithm. To obtain the probability model of a class variable C, NB uses Bayes' theorem:

$$p(C|X_1, ..., Xn) = \frac{p(C)p(X_1, ..., Xn|C)}{p(X_1, ..., Xn)}, \tag{6.13}$$

and it assumes that all feature variables, $X_1, ..., Xn$ are independent. The numerator of Eq. (6.13) is equivalent to the joint probability model, $p(C, X_1, ..., Xn)$, which can be expressed as $p(C)\prod_{i=1}^{n} p(X_i|C)$ by the definition of conditionally probability when $X_1, ..., Xn$ are independent. Since the denominator of Eq. (6.13) can be considered as constant, the probability model for a classifier is $p(C|X_1, ..., Xn) \propto p(C)\prod_{i=1}^{n} p(X_i|C)$. It is easy to calculate the probabilities for the classes from training with this model especially when it is log transformed. The assumption of independence of feature variables is almost always wrong, therefore NB classifiers are usually considered as less accurate (165). Nonetheless, despite its naive and over simplified assumptions, NB classifiers outperform other sophisticated learning algorithms in some cases (184, 191).

k-NN is an instance-based learning algorithm, which delays the induction or generalization process until the classification phase (165). When a data point needs to be classified, the distances of the data point from all the training data points are calculated. Subsequently, all the training points are sorted by the calculated distance, and the label that is the most frequent in top k points is selected as output (157). k-NN is very simple, and it can use any type of distance metrics, such as Euclidean, Manhattan, and Minkowsky (165). However, it is computationally inefficient in some cases especially when the data size is large because k-NN requires to calculate the lengths at each time of classification.

Both SVM and ANN tend to perform better on data with multi-dimensions and continuous features, but they require large sample size to achieve high accuracy. NB works well with relatively small sample size, but it requires strong (naive) independence assumptions. Decision tree learning performs well with classifying categorical data (165). It is usually based on a heuristic algorithm, and it fails to find optimal solutions in some cases. k-NN is very simple to understand and interpret, but it is sensitive to irrelevant features (165).

7

Computational implementation

This chapter gives an overview of the computational implementation approaches used in our research. We mainly used Python for general programming, and R for statistical programming and analysis. This chapter also briefly gives additional information about software development methodologies and data storage.

7.1 Software development methodologies: Rapid application development and Test-driven development

Many conventional software development methodologies, such as variants of the waterfall method (192), require strict documentation at the design phase, and full implementation rather than prototyping in the implementation phase. A waterfall method is a sequential design process flowing steadily downwards through the phases of the software development life cycle. These strict and non-flexible software development phases of conventional methodologies led to many failed system development projects in the 1980s and early 1990s, especially in large-scale projects. Rapid application development (RAD) (193) is a relatively new software development methodology, which aims to decrease the complexity of implementation and increase the speed of application development. There are many types of RADs, such as Scrum (194), Agile software development (Agile) (195), and Extreme Programming (XP) (196). Most of them focus on simplifying each phase and reducing the duration of the software development cycle. One of the most suitable RADs for small or mid-size bioinformatics projects is Test-driven development (TDD) (197), which is equivalent to the test-first programming

concepts of XP. TDD enforces the creation of unit tests before actual coding. Although full compliance to TDD is not necessary, making as many unit tests as possible with mock data can ensure high reliability of the application.

7.2 Programming languages: Object-oriented programming and Python

The most popular programming paradigm today is object-oriented programming (OOP). OOP uses "objects" defined by corresponding "classes". Objects are actual data with data structure and procedures, whereas classes are definitions or templates to make objects. Many programming languages currently support full or partial OOP, and some of the popular ones are C++, Java, Perl, and Python. All of these programming languages are freely available (Table 7.1), and they are widely used in bioinformatics analysis. Table 7.1 also shows two additional languages Haskell and Go, as examples of functional programming and concurrent programming paradigms, though these paradigms are less common than OOP in general.

Table 7.1: Programming languages - The table shows a list of freely available programming languages. "program", "type", and "URL for software environment" represent the name of programming language, the name of programming paradigm, and the URL for downloading software environment, respectively. OOP, Func, Conc in the "type" column represents three different programming paradigms: object-oriented, functional, and concurrent programming, respectively.

Program	Type	URL for software environment
C++ (GNU)	OOP	`http://gcc.gnu.org`
Java (Oracle)	OOP	`http://www.oracle.com/technetwork/java`
Perl	OOP	`http://www.perl.org`
Python	OOP	`http://www.python.org`
Haskell	Func	`http://www.haskell.org`
Go	Conc	`http://golang.org`

Among them, Python has emerged as one of the most popular languages in bioinformatics. Python requires no static type checking, which enhances the productivity with the RAD approach. It also offers multiple programming paradigms, therefore,

it can use both object-oriented and functional programming in the same module, for instance. Moreover, Python offers two very powerful libraries for scientific computing: NumPy (`http://numpy.scipy.org`) and SciPy (`http://www.scipy.org`). BioPython (`http://biopython.org`) provides a set of libraries for biological computation to Python. A machine learning package for Python called PyML provides useful functions to build and test a SVM model (175). We mainly used PyML to build our two-step SVM model for miRNA target prediction.

7.3 Statistical programming languages: R and other statistical computing languages

Some programming languages are specialized for statistical computing and graphics. For instance, SAS, SPSS, STATA, and R support software environments for statistical computing, whereas MATLAB and Mathematica are languages that provide statistical features. Among them, only R is open source software and freely available (Table 7.2). Moreover, R provides many additional libraries and also a comprehensive framework for high-throughput genome data analysis, called Bioconductor (`http://www.bioconductor.org`), hence, it is the most popular statistical computing language for bioinformatics today.

Table 7.2: Programming languages for statistical analysis - The table shows a list of programming languages that can be used for statistical analysis. "program", license", and "URL" represent the name of programming language, the type of license, and the URL for organizations or institutes that provide the software, respectively.

Program	License	URL
SAS	proprietary	`http://www.sas.com`
SPSS	proprietary	`http://www.spss.com`
STATA	proprietary	`http://www.stata.com`
R	open source	`http://www.r-project.org`
MATLAB	proprietary	`http://www.mathworks.com/products/matlab`
Mathematica	proprietary	`http://www.wolfram.com/mathematica`

7.4 Data storage: Text files and MySQL

Handling the large size data from experiments with high-throughput technologies usually requires a method for effective data retrieval and manipulation. The easiest approach is using a text file with user-defined or pre-defined format. Some examples of popular pre-defined formats in bioinformatics are FASTA for nucleotide and peptide sequences, GFF (general feature format) for positional information with additional features in genome, and MAF (multiple alignment format) for multiple alignments.

A relational database (RDB) (198) management system offers more elaborate data storage mechanisms than simple text files. In RDB, data are usually accessed through the structured query language (SQL), and all data are stored in multiple tables. MySQL is the most popular freely available RDB used by bioinformatics projects, but other free RDBs, such as PostgreSQL or Postgres, and SQLite, are also widely used (Table 7.3).

Table 7.3: Relational databases - The table shows a list of relational and NoSQL databases. [1]PostgreSQL Global Development Group. [2]SQLite Consortium. [3]Google App Engine provides BigTable accessibility.

Name	Type	Provider	URL
MySQL	RDB	Oracle	`http://www.mysql.com`
PostgreSQL	RDB	PGDG[1]	`http://www.postgresql.org`
SQLite	RDB	SQLite Cons[2]	`http://www.sqlite.org`
MongoDB	NoSQL	10gen	`http://www.mongodb.org`
BigTable	NoSQL	Google	`http://code.google.com/appengine`[3]
SimpleDB	NoSQL	Amazon	`http://aws.amazon.com/simpledb`

In RDB, the data in multiple tables are "joined" when retrieving them together. All RDB management systems have very poor performance with joining tables at tera- or peta- byte level. Therefore, NoSQL database management systems have emerged to control data even at peta byte level. NoSQL usually avoids SQL usage and relational tables. Many NoSQL systems are available today, and some popular NoSQLs are MongoDB, Google BigTable (199), and Amazon SimpleDB (Table 7.3).

Even though handling data at tera byte level is important as more sequence data

from the next generation sequencing become available, many programming languages currently lack easy-to-use libraries to access NoSQL management systems. Therefore, using both text files and RDBs rather than NoSQL is still a major practice in bioinformatics.

8

Future perspectives

This chapter gives potential improvements in context of the three sub-goals of our research. Many further perspectives are covered in the papers, therefore, this section covers only additional perspectives that are not discussed in the papers.

Firstly, for *miRNA target prediction*, it is important to improve the computational speed of our two-step SVM model at the classification phase. A current computational limitation of our model at classification is mainly caused due to its usage of non-linear kernel at the second or global level. Therefore, optimizing the second level classifier with a linear kernel is a simple solution to improve the computational speed at the classification phase. Moreover, training a SVM model with AGO pull-down data is simple, but it can be a very effective approach. However, existing AGO pull-down experiments provide no information regarding difference of miRNA binding sites between controls and transfected samples, therefore some scores, such as log-ratio values in microarray, need to be calculated for SVM training.

Secondly, the results from our study in *miRNA high-throughput experiments* can potentially improve other studies in both *miRNA target prediction* and *miRNA and other ncRNAs* because they mainly rely on the data from microarray, proteomics, and next generation sequencing for their model building and statistical analysis. Further interesting work can be an expansion of the analysis with different types of data, such as expression profiles of NCI-60 microarray data sets or time points data.

Thirdly, understanding of ncRNAs characteristics and their interactions with miRNAs becomes more important because ncRNA:ncRNA interactions are likely involved

in many gene regulations. In addition to cis-NATs and CARs, it is interesting to expand the same approach to other types of ncRNAs, such as long ncRNAs.

In conclusion, using data from next generation sequencing as well as considering the results from *miRNA high-throughput experiments* is most likely to enhance other studies in both *miRNA target prediction* and *miRNA and other ncRNAs*.

References

[1] Orgel LE, Crick FHC (1980) **Selfish DNA: the ultimate parasite**. Nature 284:604–607. 1

[2] Wright MW, Bruford EA (2011) **Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature**. Human Genomics 5:90. 1

[3] Bartel DP (2004) **MicroRNAs: genomics, biogenesis, mechanism, and function**. Cell 116:281–297. 1, 9, 10, 12, 13, 18

[4] Bartel DP (2009) **MicroRNAs: target recognition and regulatory functions**. Cell 136:215–233. 1, 15, 16, 17, 18

[5] Friedman RC, Farh KKH, Burge CB, Bartel DP (2008) **Most mammalian mRNAs are conserved targets of microRNAs**. Genome Research 19:92–105. 1, 15, 16

[6] Wang L, Oberg AL, Asmann YW, Sicotte H, McDonnell SK, et al. (2009) **Genome-wide transcriptional profiling reveals MicroRNA-correlated genes and biological processes in human lymphoblastoid cell lines**. PLoS ONE 4:e5878. 1

[7] Baek D, Villén J, Shin C, Camargo FD, Gygi SP, et al. (2008) **The impact of microRNAs on protein output**. Nature 455:64–71. 2

[8] Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, et al. (2008) **Widespread changes in protein synthesis induced by microRNAs**. Nature 455:58–63. 2

[9] Wen J, Parker BJ, Jacobsen A, Krogh A (2011) **MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action**. RNA 17:820–834. 2

[10] Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R (2008) **MicroRNA-373 induces expression of genes with complementary promoter sequences**. Proceedings of the National Academy of Sciences 105:1608–1613. 2, 18

[11] Kim DH, Saetrom P, Snove O, Rossi JJ (2008) **MicroRNA-directed transcriptional gene silencing in mammalian cells**. Proceedings of the National Academy of Sciences 105:16230–16235. 2, 18

[12] Younger ST, Corey DR (2011) **Transcriptional gene silencing in mammalian cells by miRNA mimics that target gene promoters**. Nucleic Acids Research 39:5682–5691. 2

[13] Han J, Kim D, Morris KV (2007) **Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells**. Proceedings of the National Academy of Sciences 104:12422–12427. 2

[14] Morris KV (2011) **The emerging role of RNA in the regulation of gene transcription in human cells**. Seminars in Cell & Developmental Biology 22:351–358. 2

[15] Morris KV, Santoso S, Turner AM, Pastori C, Hawkins PG (2008) **Bidirectional transcription directs both transcriptional gene activation and suppression in human cells**. PLoS Genetics 4:e1000258. 2

[16] Schwartz JC, Younger ST, Nguyen NB, Hardy DB, Monia BP, et al. (2008) **Antisense transcripts are targets for activating small RNAs**. Nature Structural & Molecular Biology 15:842–848. 2

[17] Yue X, Schwartz JC, Chu Y, Younger ST, Gagnon KT, et al. (2010) **Transcriptional regulation by small RNAs at sequences downstream from 3′ gene termini**. Nature Chemical Biology 6:621–629. 2

[18] Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, et al. (2006) **Complex loci in human and mouse genomes**. PLoS Genetics 2:e47. 3, 18, 19, 20

[19] Lee RC, Feinbaum RL, Ambros V (1993) **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. Cell 75:843–854. 9

[20] Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, et al. (1998) **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. Nature 391:806–811. 10

[21] Hamilton AJ, Baulcombe DC (1999) **A species of small antisense RNA in posttranscriptional gene silencing in plants**. Science 286:950–952. 10

[22] Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, et al. (2001) **Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells**. Nature 411:494–498. 10

[23] Mello CC, Conte D (2004) **Revealing the world of RNA interference**. Nature 431:338–342. 10, 12, 13

[24] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, et al. (2000) **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans**. Nature 403:901–906. 10

[25] Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, et al. (2000) **Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA**. Nature 408:86–89. 10

[26] Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) **Identification of novel genes coding for small expressed RNAs**. Science 294:853–858. 10

[27] Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) **An abundant class of tiny RNAs with probable regulatory roles in caenorhabditis elegans**. Science 294:858–862. 10, 12

REFERENCES

[28] Lee RC, Ambros V (2001) **An extensive class of small RNAs in Caenorhabditis elegans**. Science 294:862–864. 10

[29] Griffiths-Jones S (2004) **The microRNA registry**. Nucleic Acids Research 32:109D–111. 10

[30] Griffiths-Jones S (2006) **miRBase: the MicroRNA sequence database**. In: MicroRNA Protocols, Humana Press, volume 342. pp. 129–138. doi:10.1385/1-59745-123-1:129. URL `http://dx.doi.org/10.1385/1-59745-123-1:129`. 10

[31] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2007) **miRBase: tools for microRNA genomics**. Nucleic Acids Research 36:D154–D158. 10

[32] Kozomara A, Griffiths-Jones S (2010) **miRBase: integrating microRNA annotation and deep-sequencing data**. Nucleic Acids Research 39:D152–D157. 10

[33] Nair V, Zavolan M (2006) **Virus-encoded microRNAs: novel regulators of gene expression**. Trends in Microbiology 14:169–175. 10, 15

[34] Boss IW, Plaisance KB, Renne R (2009) **Role of virus-encoded microRNAs in herpesvirus biology**. Trends in Microbiology 17:544–553. 10, 15

[35] Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) **MicroRNA genes are transcribed by RNA polymerase II**. The EMBO Journal 23:4051–4060. 10

[36] Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) **MicroRNA maturation: stepwise processing and subcellular localization**. The EMBO Journal 21:4663–4670. 10

[37] Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) **The nuclear RNase III Drosha initiates microRNA processing**. Nature 425:415–419. 12

[38] Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) **Processing of primary microRNAs by the Microprocessor complex**. Nature 432:231–235. 12

[39] Gregory RI, ping Yan K, Amuthan G, Chendrimada T, Doratotaj B, et al. (2004) **The Microprocessor complex mediates the genesis of microRNAs**. Nature 432:235–240. 12

[40] Han J, Lee Y, Yeom KH, Kim YK, Jin H, et al. (2004) **The Drosha-DGCR8 complex in primary microRNA processing**. Genes & Development 18:3016–3027. 12

[41] Landthaler M, Yalcin A, Tuschl T (2004) **The human DiGeorge syndrome critical region gene 8 and its d. Melanogaster homolog are required for miRNA biogenesis**. Current Biology 14:2162–2167. 12

[42] Sætrom P, Snøve O, Nedland M, Grünfeld TB, Lin Y, et al. (2006) **Conserved MicroRNA characteristics in mammals**. Oligonucleotides 16:115–144. 12

[43] Zeng Y, Yi R, Cullen BR (2003) **MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms**. Proceedings of the National Academy of Sciences 100:9779–9784. 12

[44] Batuwita R, Palade V (2009) **microPred: effective classification of pre-miRNAs for human miRNA gene prediction**. Bioinformatics 25:989–995. 12

[45] Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U (2004) **Nuclear export of MicroRNA precursors**. Science 303:95–98. 12

[46] Yi R, Qin Y, Macara IG, Cullen BR (2003) **Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs**. Genes & Development 17:3011–3016. 12

[47] Tomari Y, Zamore PD (2005) **Perspective: machines for RNAi**. Genes & Development 19:517–529. 12

[48] Kim VN (2005) **MicroRNA biogenesis: coordinated cropping and dicing**. Nature Reviews Molecular Cell Biology 6:376–385. 12

[49] Fagard M, Boutet S, Morel JB, Bellini C, Vaucheret H (2000) **AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals**. Proceedings of the National Academy of Sciences 97:11650–11654. 12

[50] Carthew RW, Sontheimer EJ (2009) **Origins and mechanisms of miRNAs and siRNAs**. Cell 136:642–655. 12, 13

[51] Hammond SM, Bernstein E, Beach D, Hannon GJ (2000) **An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells**. Nature 404:293–296. 12

[52] Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) **The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila**. Cell 130:89–100. 12

[53] Ruby JG, Jan CH, Bartel DP (2007) **Intronic microRNA precursors that bypass Drosha processing**. Nature 448:83–86. 12

[54] Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) **Mammalian mirtron genes**. Molecular Cell 28:328–336. 12

[55] Borchert GM, Lanier W, Davidson BL (2006) **RNA polymerase III transcribes human microRNAs**. Nature Structural & Molecular Biology 13:1097–1101. 12

[56] Batzer MA, Deininger PL (2002) **Alu repeats and human genomic diversity**. Nature Reviews Genetics 3:370–379. 12

[57] Paddison PJ, Caudy AA, Bernstein E, Hannon GJ, Conklin DS (2002) **Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells**. Genes & Development 16:948–958. 13

[58] Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) **Mammalian microRNAs predominantly act to decrease target mRNA levels**. Nature 466:835–840. 13

[59] Behm-Ansmant I, Rehwinkel J, Doerks T, Stark A, Bork P, et al. (2006) **mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes**. Genes & Development 20:1885–1898. 13, 18

[60] Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Dongen SV, et al. (2006) **Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs**. Science 312:75–79. 13, 18

[61] Wu L, Fan J, Belasco JG (2006) **MicroRNAs direct rapid deadenylation of mRNA**. Proceedings of the National Academy of Sciences 103:4034–4039. 13, 18

[62] Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, et al. (2003) **Expression profiling reveals off-target gene regulation by RNAi**. Nature Biotechnology 21:635–637. 13

[63] Stefani G, Slack FJ (2008) **Small non-coding RNAs in animal development**. Nature Reviews Molecular Cell Biology 9:219–230. 15

[64] Alvarez-Garcia I, Miska EA (2005) **MicroRNA functions in animal development and human disease**. Development 132:4653–4662. 15

[65] He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, et al. (2005) **A microRNA polycistron as a potential human oncogene**. Nature 435:828–833. 15

[66] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, et al. (2005) **MicroRNA expression profiles classify human cancers**. Nature 435:834–838. 15

[67] Chen JF, Murchison EP, Tang R, Callis TE, Tatsuguchi M, et al. (2008) **Targeted deletion of Dicer in the heart leads to dilated cardiomyopathy and heart failure**. Proceedings of the National Academy of Sciences 105:2111–2116. 15

[68] Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, et al. (2007) **Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2**. Cell 129:303–317. 15

[69] Maes O, Chertkow H, Wang E, Schipper H (2009) **MicroRNA: implications for alzheimer disease and other human CNS disorders**. Current Genomics 10:154–168. 15

[70] Lewis BP, Shih Ih, Jones-Rhoades MW, Bartel DP, Burge CB (2003) **Prediction of mammalian MicroRNA targets**. Cell 115:787–798. 15

[71] Brennecke J, Stark A, Russell RB, Cohen SM (2005) **Principles of MicroRNA-target recognition**. PLoS Biology 3:e85. 15

[72] Lewis BP, Burge CB, Bartel DP (2005) **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets**. Cell 120:15–20. 15

[73] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M (2007) **Inference of miRNA targets using evolutionary conservation and pathway analysis**. BMC Bioinformatics 8:69. 15

[74] Ellwanger DC, Büttner FA, Mewes HW, Stümpflen V (2011) **The sufficient minimal set of miRNA seed types**. Bioinformatics 27:1346–1350. 15

[75] Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007) **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. Molecular Cell 27:91–105. 16, 17

[76] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, et al. (2005) **Combinatorial microRNA target predictions**. Nature Genetics 37:495–500. 16

[77] Enright AJ, John B, Gaul U, Tuschl T, Sander C, et al. (2003) **MicroRNA targets in Drosophila**. Genome Biology 5:R1. 16

[78] Lall S, Grün D, Krek A, Chen K, Wang YL, et al. (2006) **A genome-wide map of conserved MicroRNA targets in C. Elegans**. Current Biology 16:460–471. 16

[79] Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. Genome Biology 11:R90. 16

[80] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) **The role of site accessibility in microRNA target recognition**. Nature Genetics 39:1278–1284. 16, 17

[81] Sethupathy P, Megraw M, Hatzigeorgiou AG (2006) **A guide through present computational approaches for the identification of mammalian microRNA targets**. Nature Methods 3:881–886. 16

[82] Long D, Lee R, Williams P, Chan CY, Ambros V, et al. (2007) **Potent effect of target structure on microRNA function**. Nature Structural & Molecular Biology 14:287–294. 17

[83] John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. (2004) **Human MicroRNA targets**. PLoS Biology 2:e363. 17

[84] Saetrom P, Heale BSE, Snøve O, Aagaard L, Alluin J, et al. (2007) **Distance constraints between microRNA target sites dictate efficacy and cooperativity.** Nucleic Acids Res 35:2333–2342. 17

[85] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. (2010) **Transcriptome-wide identification of RNA-binding protein and MicroRNA target sites by PAR-CLIP**. Cell 141:129–141. 18

[86] Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, et al. (2009) **New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites**. Genome Research 19:1175–1183. 18

[87] Gu S, Jin L, Zhang F, Sarnow P, Kay MA (2009) **Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs**. Nature Structural & Molecular Biology 16:144–150. 18

[88] Duursma AM, Kedde M, Schrier M, le Sage C, Agami R (2008) **miR-148 targets human DNMT3b protein coding region**. RNA 14:872–877. 18

REFERENCES

[89] Elcheva I, Goswami S, Noubissi FK, Spiegelman VS (2009) **CRD-BP protects the coding region of βTrCP1 mRNA from miR-183-mediated degradation**. Molecular Cell 35:240–246. 18

[90] Forman JJ, Legesse-Miller A, Coller HA (2008) **A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence**. Proceedings of the National Academy of Sciences 105:14879–14884. 18

[91] Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I (2008) **MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation**. Nature 455:1124–1128. 18

[92] Hwang HW, Wentzel EA, Mendell JT (2007) **A hexanucleotide element directs MicroRNA nuclear import**. Science 315:97–100. 18

[93] Politz JCR, Zhang F, Pederson T (2006) **MicroRNA-206 colocalizes with ribosome-rich regions in both the nucleolus and cytoplasm of rat myogenic cells**. Proceedings of the National Academy of Sciences 103:18957–18962. 18

[94] Duhig T, Ruhrberg C, Mor O, Fried M (1998) **The human surfeit locus**. Genomics 52:72–78. 19

[95] Holmes R, Williamson C, Peters J, Denny P, Group RIKENGER, et al. (2003) **A comprehensive transcript map of the mouse Gnas imprinted complex**. Genome Research 13:1410–1415. 19

[96] Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, et al. (2003) **An abundance of bidirectional promoters in the human genome**. Genome Research 14:62–66. 19, 20

[97] Wang XJ, Gaasterland T, Chua NH (2005) **Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana.** Genome Biology 6:R30. 19

[98] Chen J, Sun M, Kent WJ, Huang X, Xie H, et al. (2004) **Over 20transcripts might form sense-antisense pairs**. Nucleic Acids Research 32:4812–4820. 19

[99] Zhang Y, Liu XS, Liu QR, Wei L (2006) **Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species**. Nucleic Acids Research 34:3465–3475. 19

[100] Osato N, Suzuki Y, Ikeo K, Gojobori T (2007) **Transcriptional interferences in cis natural antisense transcripts of humans and mice**. Genetics 176:1299–1306. 19, 20

[101] Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005) **Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in arabidopsis**. Cell 123:1279–1291. 20

[102] Tufarelli C, Stanley JAS, Garrick D, Sharpe JA, Ayyub H, et al. (2003) **Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease**. Nature Genetics 34:157–165. 20

[103] Beisel C, Paro R (2011) **Silencing chromatin: comparing modes and mechanisms**. Nature Reviews Genetics 12:123–135. 20, 21

[104] Mondal T, Rasmussen M, Pandey GK, Isaksson A, Kanduri C (2010) **Characterization of the RNA content of chromatin**. Genome Research 20:899–907. 20

[105] Mercer TR, Dinger ME, Mattick JS (2009) **Long noncoding RNAs: insights into functions**. Nature Reviews Genetics 10:155–159. 21

[106] Whitehead J, Pandey GK, Kanduri C (2009) **Regulation of the mammalian epigenome by long noncoding RNAs**. Biochimica et Biophysica Acta (BBA) - General Subjects 1790:936–947. 21

[107] Alwine JC, Kemp DJ, Stark GR (1977) **Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.** Proceedings of the National Academy of Sciences 74:5350–5354. 23

[108] Becker-André M, Hahlbrock K (1989) **Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by aPCR aidedtranscipttitration assay (PATTY)**. Nucleic Acids Research 17:9437–9446. 23

[109] Noonan KE, Beck C, Holzmayer TA, Chin JE, Wunder JS, et al. (1990) **Quantitative analysis of MDR1 (multidrug resistance) gene expression in human tumors by polymerase chain reaction.** Proceedings of the National Academy of Sciences 87:7160–7164. 23

[110] Morozova O, Hirst M, Marra MA (2009) **Applications of new sequencing technologies for transcriptome analysis**. Annual Review of Genomics and Human Genetics 10:135–151. 23, 25, 26, 27

[111] VanGuilder HD, Vrana KE, Freeman WM (2008) **Twenty-five years of quantitative PCR for gene expression analysis**. BioTechniques 44:619–626. 23

[112] Quackenbush J (2006) **Microarray analysis and tumor classification**. New England Journal of Medicine 354:2463–2472. 24

[113] Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, et al. (2006) **Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project**. Nature Biotechnology 24:1140–1150. 24

[114] Quackenbush J (2002) **Microarray data normalization and transformation**. Nature Genetics 32:496–501. 24

[115] Mockler TC, Ecker JR (2005) **Applications of DNA tiling arrays for whole-genome analysis**. Genomics 85:1–15. 25

[116] Sanger F, Nicklen S, Coulson AR (1977) **DNA sequencing with chain-terminating inhibitors**. Proceedings of the National Academy of Sciences 74:5463–5467. 25

[117] Adams M, Kelley J, Gocayne J, Dubnick M, Polymeropoulos M, et al. (1991) **Complementary DNA sequencing: expressed sequence tags and human genome project**. Science 252:1651–1656. 25

[118] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) **Serial analysis of gene expression**. Science 270:484–487. 25

[119] Voelkerding KV, Dames S, Durtschi JD (2010) **Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy**. The Journal of Molecular Diagnostics 12:539–551. 25

[120] Hall N (2007) **Advanced sequencing technologies and their wider impact in microbiology**. Journal of Experimental Biology 210:1518–1525. 26

[121] Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, et al. (2008) **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing**. BioTechniques 45:81–94. 27

[122] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) **The transcriptional landscape of the yeast genome defined by RNA sequencing**. Science 320:1344–1349. 27

[123] Sanford JR, Wang X, Mort M, VanDuyn N, Cooper DN, et al. (2008) **Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts**. Genome Research 19:381–394. 27

[124] Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, et al. (2008) **HITS-CLIP yields genome-wide insights into brain alternative RNA processing**. Nature 456:464–469. 27

[125] Jensen KB, Darnell RB (2008) **CLIP: crosslinking and ImmunoPrecipitation of in vivo RNA targets of RNA-binding proteins**. In: Methods in Molecular Biology, Humana Press, volume 488. pp. 85–98. doi:10.1007/978-1-60327-475-3_6. URL http://dx.doi.org/10.1007/978-1-60327-475-3_6. 27

[126] Wong E, Wei CL (2009) **ChIP'ing the mammalian genome: technical advances and insights into functional elements**. Genome Medicine 1:89. 27

[127] Niranjanakumari S, Lasda E, Brazas R, Garcia-Blanco MA (2002) **Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo**. Methods 26:182–190. 27

[128] Conrad NK (2008) **Co-immunoprecipitation techniques for assessing RNA-protein interactions in vivo**. In: Methods in Enzymology, Elsevier, volume 449. pp. 317–342. doi:10.1016/s0076-6879(08)02415-4. URL http://dx.doi.org/10.1016/S0076-6879(08)02415-4. 27

[129] Mili S, Steitz JA (2004) **Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses**. RNA 10:1692–1694. 27

[130] Boggess B (2001) **Mass spectrometry desk reference (Sparkman, O. David)**. Journal of Chemical Education 78:168. 28

[131] Mallick P, Kuster B (2010) **Proteomics: a pragmatic perspective**. Nature Biotechnology 28:695–709. 28

[132] Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. (2002) **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. Molecular & Cellular Proteomics 1:376–386. 28

[133] Ong SE, Mann M (2005) **Mass spectrometry-based proteomics turns quantitative**. Nature Chemical Biology 1:252–262. 28

[134] Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) **Quantitative mass spectrometry in proteomics: a critical review**. Analytical and Bioanalytical Chemistry 389:1017–1031. 28

[135] Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, et al. (2010) **ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments**. Nucleic Acids Research 39:D1002–D1004. 29

[136] Barrett T, Edgar R (2006) **[19] Gene expression omnibus: microarray data storage, submission, retrieval, and analysis**. In: Methods in Enzymology, Elsevier, volume 411. pp. 352–369. doi:10.1016/s0076-6879(06)11019-8. URL http://dx.doi.org/10.1016/S0076-6879(06)11019-8. 29

[137] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) **Minimum information about a microarray experiment (MIAME)—toward standards for microarray data**. Nature Genetics 29:365–371. 29

[138] Leinonen R, Sugawara H, Shumway M (2010) **The sequence read archive**. Nucleic Acids Research 39:D19–D21. 29

[139] Mead JA, Bianco L, Bessant C (2009) **Recent developments in public proteomic MS repositories and pipelines**. PROTEOMICS 9:861–881. 29

[140] Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, et al. (2006) **PRIDE: a public repository of protein and peptide identifications for the proteomics community**. Nucleic Acids Research 34:D659–D663. 29

[141] Craig R, Cortens JP, Beavis RC (2004) **Open source system for analyzing, validating, and storing protein identification data**. Journal of Proteome Research 3:1234–1242. 29

[142] Deutsch EW (2009) **The PeptideAtlas project**. In: Methods in Molecular Biology, Humana Press, volume 604. pp. 285–296. doi:10.1007/978-1-60761-444-9_19. URL http://dx.doi.org/10.1007/978-1-60761-444-9_19. 29

[143] Rosner B (2006) *Fundamentals of biostatistics*. Duxbury Press. 31, 32, 33, 36, 37, 38

[144] Mann HB, Whitney DR (1947) **On a test of whether one of two random variables is stochastically larger than the other**. The Annals of Mathematical Statistics 18:50–60. 33

[145] Wilcoxon F (1945) **Individual comparisons by ranking methods**. Biometrics Bulletin 1:80. 33

REFERENCES

[146] Kvam PH (2007) *Nonparametric statistics with applications to science and engineering (Wiley series in probability and statistics*. Wiley-Interscience. 34

[147] Arnold S (2005) **Nonparametric statistics**. Technical report, Penn State University. URL `http://astrostatisitics.psu-edu`. 34

[148] Marsaglia G, Tsang WW, Wang J (2003) **Evaluating Kolmogorov's distribution**. Journal of Statistical Software 8:1–4. 34

[149] Rizzo ML (2008) *Statistical computing with R*. Chapman and Hall/CRC. 34

[150] Moore DS, McCabe GP, Craig BA (2009) *Introduction to the practice of statistics*. Introduction to the Practice of Statistics. W.H. Freeman. 34, 35, 36

[151] Efron B, Tibshirani RJ (1994) *An introduction to the bootstrap (Chapman & Hall/CRC monographs on statistics & applied probability)*. Chapman and Hall/CRC, first edition. 35

[152] Fisher RA (1922) **On the interpretation of χ2 from contingency tables, and the calculation of p**. Journal of the Royal Statistical Society 85:87. 35

[153] Kruskal WH, Wallis WA (1952) **Use of ranks in one-criterion variance analysis**. Journal of the American Statistical Association 47:583–621. 36

[154] Benjamini Y, Hochberg Y (1995) **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. Journal of the Royal Statistical Society: Series B (Methodological) 57:289–300. 36

[155] Schneider A, Hommel G, Blettner M (2010) **Linear regression analysis**. Deutsches Aerzteblatt Online 107:776–782. 37

[156] Pavlidis P, Wapinski I, Noble WS (2004) **Support vector machine classification on the web**. Bioinformatics 20:586–587. 39

[157] Tarca AL, Carey VJ, Chen Xw, Romero R, Drăghici S (2007) **Machine learning and its applications to biology**. PLoS Computational Biology 3:e116. 39, 40, 45, 52, 53

[158] Boser BE, Guyon IM, Vapnik VN (1992) **A training algorithm for optimal margin classifiers**. In: Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. ACM Press, pp. 144–152. doi:10.1145/130385.130401. URL `http://dx.doi.org/10.1145/130385.130401`. 40

[159] Vapnik VN (1982) *Estimation dependences based on empirical data*. New York, USA: Springer-Verlag. 40, 41

[160] Vapnik VN (1998) *Statistical learning theory*. Wiley, New York. 40

[161] Joachims T (2002) *Learning to classify text using support vector machines*. Kluwer. 40, 41, 43, 50, 51, 52

[162] Burges CJC (1998) **A tutorial on support vector machines for pattern recognition**. Data Mining and Knowledge Discovery 2:121–167. 40

[163] Smola AJ, Schölkopf B (2003) **A tutorial on support vector regression**. Technical report, STATISTICS AND COMPUTING. 43, 51, 52

[164] Cortes C, Vapnik V (1995) **Support-vector networks**. In: Machine Learning. Springer Science and Business Media LLC, volume 20, pp. 273–297. doi:10.1007/bf00994018. URL `http://dx.doi.org/10.1007/bf00994018`. 43

[165] Kotsiantis SB (2007) **Supervised machine learning: a review of classification techniques**. Informatica 31:249–268. 43, 52, 53

[166] Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) **New support vector algorithms**. Neural Computation 12:1207–1245. 45

[167] Provost FJ, Kohavi R (1998) **Glossary of terms. on applied research in machine learning**. Machine Learning 30:271–274. 45

[168] Swets J (1988) **Measuring the accuracy of diagnostic systems**. Science 240:1285–1293. 45, 48

[169] Huang J, Ling CX (2005) **Using AUC and accuracy in evaluating learning algorithms**. IEEE Transactions on Knowledge and Data Engineering 17:299–310. 48

[170] Provost FJ, Fawcett T (1997) **Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions**. In: Knowledge Discovery and Data Mining. pp. 43–48. 48

[171] Gribskov M, Robinson NL (1996) **Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching**. Computers & Chemistry 20:25–33. 48

[172] Swamidass SJ, Azencott CA, Daily K, Baldi P (2010) **A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval**. Bioinformatics 26:1348–1356. 48

[173] Provost F (2000). **Machine learning from imbalanced data sets 101**. 50

[174] Wu G, Chang EY (2003) **Class-boundary alignment for imbalanced dataset learning**. In: In ICML 2003 Workshop on Learning from Imbalanced Data Sets. pp. 49–56. 50

[175] Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) **Support vector machines and kernels for computational biology**. PLoS Computational Biology 4:e1000173. 50, 51, 57

[176] Hsu CW, Chang CC, Lin CJ (2003) *A practical guide to support vector classification*. Taipei, Taiwan. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`. 50, 51

[177] Chang CC, Lin CJ (2011) **LIBSVM: A library for support vector machines**. ACM Transactions on Intelligent Systems and Technology 2:1–27. 51

[178] Ben-Hur A, Weston J (2009) **A user's guide to support vector machines**. In: Methods in Molecular Biology, Humana Press, volume 609. pp. 223–239. doi:10.1007/978-1-60327-241-4_13. URL http://dx.doi.org/10.1007/978-1-60327-241-4_13. 51

[179] Crammer K, Singer Y (2001) **On the algorithmic implementation of multiclass kernel-based vector machines**. Journal of Machine Learning Research 2:265–292. 51

[180] Murthy SK (1998) **Automatic construction of decision trees from data: a multi-disciplinary survey**. Data Mining and Knowledge Discovery 2:345–389. 52

[181] Rumelhart DE, Hinton GE, Williams RJ (1986) *Learning internal representations by error propagation*. Cambridge, MA, USA: MIT Press, 318–362 pp. 52

[182] Cestnik B, Kononenko I, Bratko I (1987) **ASSISTANT 86: A knowledge-elicitation tool for sophisticated users**. In: Bratko I, Lavrac N, editors, European Conference on Machine Learning (ECML). Sigma Press, Wilmslow, pp. 31–45. URL https://dl.acm.org/doi/abs/10.5555/3108739.3108742. 52

[183] Cover T, Hart P (1967) **Nearest neighbor pattern classification**. IEEE Transactions on Information Theory 13:21–27. 52

[184] Caruana R, Niculescu-Mizil A (2006) **An empirical comparison of supervised learning algorithms**. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. ACM Press, pp. 161–168. doi:10.1145/1143844.1143865. URL http://dx.doi.org/10.1145/1143844.1143865. 52, 53

[185] Yan X, Chao T, Tu K, Zhang Y, Xie L, et al. (2007) **Improving the prediction of human microRNA target genes by using ensemble algorithm**. FEBS Letters 581:1587–1593. 52

[186] Kingsford C, Salzberg SL (2008) **What are decision trees?** Nature Biotechnology 26:1011–1013. 52

[187] Breiman L (2001) **Random forests**. In: Machine Learning. Springer Science and Business Media LLC, volume 45, pp. 5–32. doi:10.1023/a:1010933404324. URL https://doi.org/10.1023/A:1010933404324. 52

[188] Krogh A (2008) **What are artificial neural networks?** Nature Biotechnology 26:195–197. 52

[189] Rosenblatt F (1957) **The perceptron: a perceiving and recognizing automaton (Project Para)**. Technical Report 85-460-1, Cornell Aeronautical Laboratory. 52

[190] Rumelhart DE, Hinton GE, Williams RJ (1986) **Learning representations by back-propagating errors**. Nature 323:533–536. 52

[191] Domingos P, Pazzani M (1997) **On the optimality of the simple bayesian classifier under zero-one loss**. Machine Learning 29:103–130. 53

[192] Royce WW, Royce WW (1970) **Managing the development of large software systems**. In: Technical Papers of Western Electronic Show and Convention. WesCon, pp. 1–9. 55

[193] Martin J (1991) *Rapid application development*. Pearson Higher Education, 736 pp. 55

[194] Rising L, Janoff NS (2000) **The Scrum software development process for small teams**. IEEE Software 17:26–32. 55

[195] Cockburn A, Highsmith J (2001) **Agile software development, the people factor**. Computer 34:131–133. 55

[196] Auer K, Miller R (2001) *Extreme programming applied: playing to win*. Addison-Wesley Professional. 55

[197] Beck K (2002) *Test driven development: by example*. Addison-Wesley Professional. 55

[198] Codd EF (1970) **A relational model of data for large shared data banks**. Communications of the ACM 13:377–387. 58

[199] Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, et al. (2008) **Bigtable: A distributed storage system for structured data**. ACM Transactions on Computer Systems 26:1–26. 58