이학석사 학위논문

가우시안 프로세스 모델과 Spherical vector clustering 기법을 이용한 시계열 데이터에서의 마이크로RNA 표적 모듈 예측

Finding Condition-specific Target Module of MicroRNA in Time Series Transcriptome Data - using Gaussian Process Model and Spherical Vector Clustering -

2018년 8월

서울대학교 대학원
협동과정 생물정보학전공
강 혜 진

가우시안 프로세스 모델과 Spherical vector clustering 기법을 이용한 시계열 데이터에서의 마이크로RNA 표적 모듈 예측

Finding Condition-specific Target Module of MicroRNA in Time Series Transcriptome Data - using Gaussian Process Model and Spherical Vector Clustering -

지도 교수 김 선

이 논문을 이학석사 학위논문으로 제출함

2018년 8월

서울대학교 대학원

협동과정 생물정보학전공

강 혜 진

강혜진의 이학석사 학위논문을 인준함

2018년 8월

위 원 장	이 병 재	(인)
부위원장	김 선	(인)
위 원	김 희 발	(인)

Abstract

Finding Condition-specific Target Module of MicroRNA in Time Series Transcriptome Data

- using Gaussian Process Model and Spherical Vector Clustering -

Kang Hyejin

Interdisciplinary Program in Bioinformatics

College of Natural Sciences

Seoul National University

MicroRNAs, widely conserved small non-coding RNAs in several species, are important key regulators which mediate post-transcriptional gene silencing. As it is known that microRNAs are involved in many important processes from cell differentiation to apoptosis in recent studies, microRNA target prediction has been studied in various ways. The most typical way for predicting

microRNA targets is to use nucleotide sequence features, which does not take into account condition-specific differences in transcript expression in cells. Therefore, a number of tools using transcript expression profiles in specific biological context have been developed to overcome the weakness of the traditional methods based on sequence features. But there are few proposed tools for time-series transcriptome dataset that provides dynamic expression patterns of microRNAs and their target mRNAs which can improve accuracy of target prediction.

In this paper, a new pipeline is proposed that predicts microRNA targets by integrating sequence feature and time-series expression profiles in specific experimental condition. For two datasets with different experimental conditions and cell types, condition specific target modules were predicted with our new pipeline for differentially expressed microRNAs that were reported from original papers. The context specificity of target modules was measured with three (correlation-based, target gene-based, network-based) evaluation criteria. MirTime showed good performance in three criteria more consistently than other microRNA target prediction methods using expression profiles.

Keyword : Gaussian process regression model, MicroRNA, Spherical vector clustering, Target prediction, Time-series data analysis, Transcriptome analysis

Student Number : 2016-20460

Contents

Abstract i

Contents iv

List of Figures vi

Chapter 1. Introduction 1

1.1. Target prediction of microRNA 1

1.2. Target prediction tools using expression profiles 2

1.3. Motivation 4

1.4. Challenges using time-series expression data 5

Chapter 2. Method and Materials 8

2.1 Sequence-based filtering 10

2.2 Condition specific target module selection using time series expression profile 10

2.2.1 Gaussian process regression model 10

2.2.2 Calculation of GP-weight vector 12

2.2.3 Spherical k-means clustering and cluster scoring 12

2.3 Evaluation criteria of context specificity of target modules 13

Chapter 3. Results 16

3.1 Time-series transcriptome datasets 16

3.2 Comparison with other tools using expression profiles 17

3.3 Performance comparison with other tools on A375 malignant melanoma cell data 18

3.3.1 Correlation-based evaluation 23

3.3.2 Target gene-based evaluation 24

3.3.3 Network-based evaluation 24

3.4 Performance comparison with other tools on MCF-7 breast cancer cell data 25

3.4.1 Correlation-based evaluation 26

3.4.2 Target gene-based evaluation 27

3.4.3 Network-based evaluation 27

Chapter 4. Discussion 30

Bibliography 32

Reference 32

요약 32

감사의 글 32

List of Figures

Figure 1. Gaussian process regression model of hsa-miR-503-5p and APLP2 in MCF-7 breast cancer cell data after training. Red dots shows the average expression value of three replicates data after z-normalization used in training, and red lines shows the max and minimum value of the three replicates. Blue shadow shows the covariances. APLP2 is one of the genes in predicted target module of hsa-miR-503-5p 7

Figure 2. Workflow. First, candidate gene groups targeted by miRNA are filtered using sequence information, previously existing target prediction methods and experimentally supported miRNA target databases. The time-series expression vectors are corrected using Gaussian process regression model only for the miRNA and genes which are selected from step 1. In step 2, after the time-series data

correction, candidate genes are grouped into clusters using Spherical k-means algorithm. The cluster showing the most repressed expression profile by the corresponding miRNA is selected as the target module 9

Figure 3. Performance comparison with other existing tools for the A375 dataset. (A) The number of target genes for 20 DEmiRNAs. The number on the right side of the bar denotes average number of target genes. (B) The mean of expression correlations between miRNAs and target genes. (C) he mean of expression correlations between target genes for each miRNA. (D) Association between target genes and genes that were reported in the original paper (n = 100). (E) Visualization of miRNA-target gene networks. The black nodes are miRNA and the yellow and green nodes are the predicted target genes by each method. The green genes are co-targeted genes by at least two miRNAs. (F) The enrichged GO term for eight modules that were provived from mirTime's network 22

Figure 4. Performace comparison with other existing tools for the MCF7 dataset. (A) The number of target genes for 9 DEmiRNAs. The number on the right side of the bar denotes average number of target genes. (B) The mean of expression

correlations between miRNAs and target genes. (C) The mean of expression correlations between target genes for each miRNA. (D) Association between target genes and genes that were reported in the original paper (n = 100). (E) Visualization of miRNA–target gene networks. The black nodes are miRNA and the yellow and green nodes are the predicted target genes by each method. The green genes are co–targeted genes by at least two miRNAs. (F) The enrichged GO term for six modules that were provived from mirTime's network 28

Chapter 1. Introduction

1.1. Target prediction of microRNA

MicroRNAs (miRNAs) are small non-coding RNAs that post-transcriptionally regulate messenger RNAs (mRNAs). A mature miRNA, incorporated into a RNA-induced silencing complex (RISC), guide RISC to help cut off the target mRNA or inhibit translation [1]. In mammalians, it is usually known that a mature miRNA has 6-8nt seed region at 5' terminal which binds to the 3 ' untranslated regions (3' UTR) of target mRNA with complementary base sequence [2]. But for the short size of the seed region, the number of false positive mRNA targets becomes very large if we use the only information that the seed region has a complementary base sequence to the target sequence. So there have been many computational methods using additional sequence features to increase the accuracy of the target prediction: complementary binding sites at 3'UTR of target mRNA,

the 3'UTR length of target mRNA, evolutionary conservation, local structure, AU content and thermodynamic binding stability, etc. PITA [3], PicTar [4], RNA22 [5] and TargetScan [6] are the representative target prediction algorithms using various sequence features. For example, TargetScan improved their target prediction algorithm with the linear regression model using site type and another 14 features.

1.2. Target prediction tools using expression profiles

The traditional target prediction tools based on nucleotide sequence features are designed to predict targets across all known transcripts, so they do not take into account differences in transcript expression in cells due to differences in cell lines or other experimental conditions. When transcripts are expressed at different concentrations, target space of miRNAs could change, thus same miRNA could act differently within particular cell types or experimental conditions. Recently, there has been new tools using transcriptome expression profiles in specific biological context for miRNA function prediction. Representative target prediction tools for this type are TargetExpress [7], MMIA [8], MAGIA2 [9], mirConnX

[10], KNN-MI [11], MINE [12] and GenMiR++ [13].

The tools using expression profiles can be further classified into three types according to the number of samples of expression data received as input. First, TargetExpress [7], using the Support Vector Machine previously trained with independent expression profiles from various cell types, receives one sample expression profiles and predicts the targets. Second, there are tools that accept input data with two classes of samples, control and treated. The representative example of this type is MMIA [8], which is a web server integrating miRNA and mRNA expression data with predicted miRNA target information from TargetScan, PITA and PicTar for analyzing miRNA-associated phenotypes and biological functions by gene set analysis. Lastly, there are tools to predict miRNA targets by receiving multiple sample data. Magia2 [9], mirConnX [10], KNN-MI[11], MINE [12], GenMiR++ [13] and methods using Pearsons correlation coefficient [14] and Spearman correlation coefficient [15] between miRNA and mRNA pairs crresponds to this type.

1.3. Motivation

Methods for miRNA target prediction so far have achieved better performance using expression profiles than traditional methods. However, integration of specific contexts like time-series experiments can provide additional information on dynamic changes in gene regulation compared to data focused on a single time point [14]. Nevertheless, there are few miRNA target prediction tools considering the characteristics of biological time series data.

A typical workflow of predicting miRNA-target in time-series data does not consider time-to-time dependency of time series dataset. For example, Lu [15] and Grilli [14] measured the mRNA and miRNA expression over a time-series in bone marrow-derived macrophages and Osteosarcoma, respectively. They identified differentially expressed miRNAs, collected putative targets from databases such as TargetScan that predict miRNA targets using sequence information, and chose miRNA-target pairs with strong negative correlation coeffiecient (Pearson's or Spearman) with time-series samples. However, these multi-sample methods treat each time-point as an independent sample, leading to a loss of information. Therefore, we aim to develop a new method that uses time domain information to predict miRNA targets in analysis of

time-series expression data.

1.4. Challenges using time-series expression data

There are two major challenges for predicting miRNA targets using time-series expression data. First one is large dimension but small-sized data. According to GENCODE [16] Version 28, 1,881 miRNAs and 19,901 protein coding genes are present in human species. However, due to the cost of experiment, the number of time points in time series data is generally small. Also, the interval between the time axes is not constant, because time-points are selectively determined according to the experimental environment and design. Second, unpaired time-points between miRNA and mRNA expression data. Since the number of miRNAs and mRNAs is very different by tens to thousands each, they might be measured separately in each experiment with unpaired time points. For example, Kreis and colleagues first measured miRNA expression in human malignant melanome cell line at 9 time points (0, 0.5, 3, 6, 12, 24, 48, 72, 96 hours) [17] and measured mRNA expression afterwards at 6 time points (0, 3, 12, 24, 48, 72 hours) to study interplay miRNA and mRNA [18].

The main idea we propose for the challenges of time series is a Gaussian process model. Modeling time series as a Gaussian process, we can predict Gaussian mean and variance of each time point including unobserved ones using information from the surrounding time-points. Figure 1 shows Gaussian process regression model of hsa-miR-503-5p (a regulator) and APLP2 (a target gene of hsa-miR-503-5p) in MCF-7 breast cancer cell data.

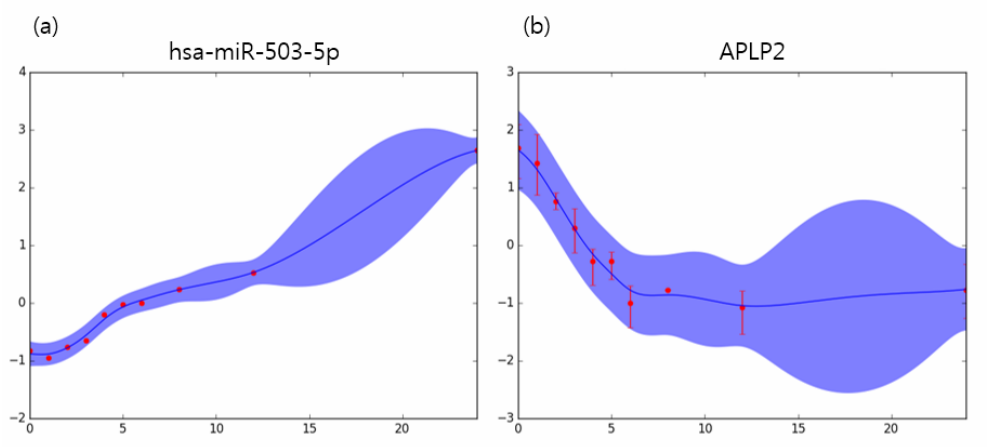


Figure 1. Gaussian process regression model of hsa-miR-503-5p and APLP2 in MCF-7 breast cancer cell data after training. Red dots shows the average expression value of three replicates data after z-normalization used in training, and red lines shows the max and minimum value of the three replicates. Blue shadow shows the covariances. APLP2 is one of the genes in predicted target module of hsa-miR-503-5p.

Chapter 2. Method and Materials

In this paper, we suggest a new miRNA target module prediction algorithm combining three kinds of information: sequence level features, expression profile along different cell lines, and time-series experiments for specific context. The overall target prediction algorithm consists of two steps as shown in the Figure 1. First the gene groups targeted by each miRNA are filtered using sequence information, previously existing target prediction methods, and experimentally supported miRNA target databases. Next, using Gaussian process regression models trained with time-series expression data and spherical K-means clustering algorithm, we extracted the target gene module for each miRNA. In Result section, miRNA target modules extracted by our pipeline was compared with other tools using expression profiles in context-specific aspects with three different evaluation criteria. The effect of time-series data correction with Gaussian process regression model has also been evaluated.

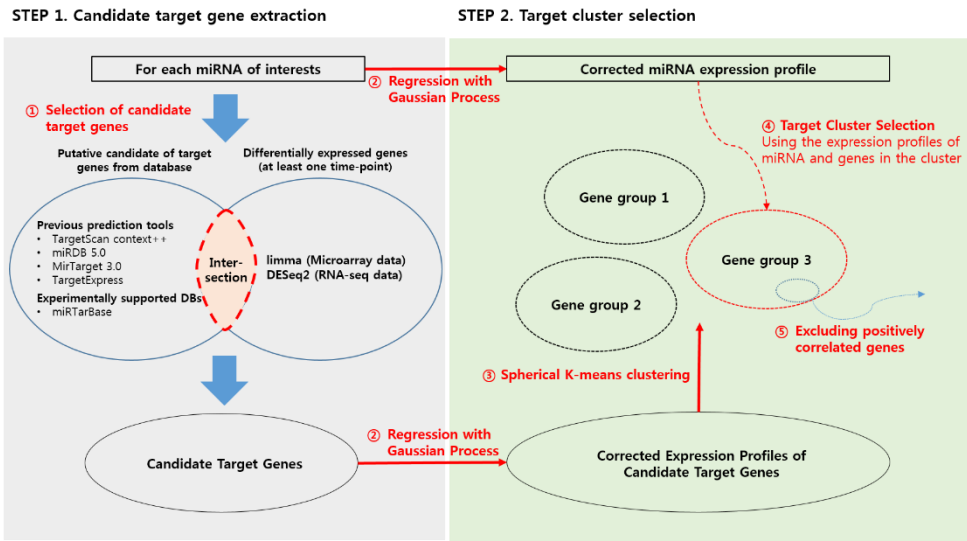


Figure 2. Workflow. First, candidate gene groups targeted by miRNA are filtered using sequence information, previously existing target prediction methods and experimentally supported miRNA target databases. The time-series expression vectors are corrected using Gaussian process regression model only for the miRNA and genes which are selected from step 1. In step 2, after the time-series data correction, candidate genes are grouped into clusters using Spherical k-means algorithm. The cluster showing the most repressed expression profile by the corresponding miRNA is selected as the target module.

2.1. Sequence-based filtering

Before using time-series expression data, we extracted candidate target gene set using previous target prediction tools and experimentally supported database. Candidate target genes of each miRNA are union of following four sets: the genes with 8-mer canonical target sites in the 3'UTR, the genes which are predicted by TargetScan 7.2 (conserved sites of conserved miRNA families only), the genes registered in miRTarBase 7.0, and the genes predicted by TargetExpress on at least one time point.

2.2. Condition specific target module selection using time-series expression data

2.2.1. Gaussian process regression model

A Gaussian process is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. The distribution of a Gaussian process is the joint distribution of infinitely many random variables, and as such, it is a

distribution over functions with a continuous domain, for example, time in our case. For Gaussian process, f(x)

$$[f_1, f_2, ..., f_k] \sim \mathcal{N}[m, K].$$

m is the mean of distribution and K is the covariance kernel which has the (i,j) element as $k(x_i, x_j)$, the kernel function calculating the covariance between x_I and x_j. In this paper we use Matern kernel (1) for the kernel k and hyperparameter ν is set 1.5, which shows the smallest root meat square error betweet real expression vector and trained model for both dataset we use. Parameter σ and l is refined during training.

$$k(x_i, x_j) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\gamma\sqrt{2\nu}d\left(\frac{x_i}{l}, \frac{x_j}{l}\right)\right)^{\nu} K_\nu \left(\gamma\sqrt{2\nu}d\left(\frac{x_i}{l}, \frac{x_j}{l}\right)\right) \quad (1)$$

Gaussian process regression model is the model that can infer the function from the probabilistic function distribution defined by training. After trained with the input data $D = \{(x_i, y_i | i = 1,2, \dots, n\}$, given the m number of test data $x_1^*, x_2^*, \dots, x_m^*$, the function $f_1^*, f_2^*, \dots, f_m^*$ is inferenced as following equation (2). When $k^* = [k(x_1, x^*, k(x_2, x^*), \dots, k(x_m, x^*)]$

$$f^*|D, x^* \sim \mathcal{N}\left(k^{*T}K^{-1}f, k(x^*, x^*) - k^{*T}K^{-1}k^*\right) \quad (2)$$

For only the miRNAs and genes selected in sequence-based filtering step, each Gaussian process regression model is trained with time-

series expression profiles. For each time point x_I, y_i is the expression value after z-normalization. Figure 1. is the example of Gaussian proceses model after training.

2.2.2. Calculation of GP-weight vector

From the trained regression model, the mean and variance vectors were estimated for each point with input values at equal intervals from the minimum time point to maximum time point given experiment data. In this case, the minimum interval between time axes is set to coincide with the minimum interval of the actual data. The Gaussian process weight vector (GP-weight vector) was calculated for each miRNA and gene with an mean and variance vector. The larger the variance value at the inferred point, the lower the reliability of the regression result. Therefore, the gpWeight vector is obtained by dividing the mean value deduced from each time axis by the variance value and used in the following analysis.

2.2.3. Spherical k-means clustering and cluster scoring

The target gene group of each miRNA based on sequence information was divided into small clusters by performing Spherical

K-means algorithm. GPweight vectors of genes were used as input values. Spherical K-means is an algorithm that generates k clusters so that the cosine distance between vectors in one cluster is minimized. For each miRNA, the share of the target cluster gene number of each miRNA divided by 100 was taken as k so that about 100 genes could be assigned per cluster.

After clustering, the Pearson correlation coefficients between the gene belonging to each cluster and the GPweight vector of the miRNA targeting the gene was calculated. If the target relationship of miRNA and gene is valid, it is assumed that there will be a negative correlation between the two expression vectors. Therefore, we calculated the average of correlation coefficient for each cluster and selected clusters with the smallest mean value and less than p-value 0.05 as effective target clusters of the miRNA.

2.3. Evaluation criteria of context specificity of target modules

To verify that the predicted target modules actually show context specificity with regard to the experimental condition, we performed three evaluation analysis on the target modules of differentially expressed miRNAs (DEmiRNAs, hereafter) reported in the reference

paper.

- Correlation-based evaluation : First, the expression correlations between miRNA and the predicted target genes are investigated. Since miRNAs suppress the expression level of the target gene, putative miRNA-target gene pairs are expected to show highly negative correlations. Second, the expression correlations between the target genes within the miRNA-target module are investigated. If a pair of target genes was targeted by the same miRNA, they will show high expression correlations. We used the Pearson's correlation coeifficient [14] between the two transcripts as a measure of the expression correlations.
- Target gene-based evaluation : The association between the predicted target genes and published analytical results for that data is investigated. When we think of the published results as a reliable analitic result, this association implies the accuracy of the target gene prediction. We conducted Fisher's exact test as a measure of the association.
- Network-based evaluation : The miRNA-target gene networks were visualized by a tool, Cytoscape [19]. From the visulaization, we investigated the topological structure of the networks. Then, a sub-modules were produced by a

network clustering method [20], then the biological functions of the modules were investigated by association test for the gene ontology biological process terms [21].

For a comparison analysis, our tool, mirTime, was compared with nine miRNA-target prediction methods using both sequence and expression: targetExpress [7], GenMir++ [13], MMIA [8], Magia2 [9], KNN-MI [11], MINE [12], Pearsons correlation coefficient (PCC) [14], and Spearman correlation coefficient (SCC) [15].

We conducted KNN-MI, MINE, PCC, and SCC using implemetation of mirTarVis [22]. Among them, targetExpress, GenMir++, and MMIA takes single sample or paired two samples (control vs. case). To apply these tools to time-series sample data, we run the tool at each time point and integrated the results by two startages: time-domain intersection (TDI) and time-domain union (TDU). Also, we added mirTime-noGP (mirTime framework without using gaussian process model) to investigate effect of gaussian process model.

Chapter 3. Result

3.1. Time-series transcriptome datasets

In this paper, the algorithm has been tested on two sets of time-series expression profiles from microarray and RNA-sequencing, respectively.

Microarray data was downloaded from ArrayExpress (E-MEXP-3544, E-MEXP-3720) [18]. It is human malignant melanome cell line data, A375, following interferon-gamma-induced gene transcription. This dataset has different number of time points for miRNA (9 time points: 0, 0.5, 3, 6, 12, 24, 48, 72, 96 hours) and mRNA (6 time points: 0, 3, 12, 24, 48, 72 hours). Both miRNA and mRNA CEL files were normalized to RMA values by R package Oligo [23] and probes with same RefSeq ID were averaged.

RNA-seq data is from the Gene Expression Ominibus (GEO). It is MCF-7 breast cancer cell data (GSE78167) responding to

estrogen following a period of estrogen starvation [24]. Three independent biological replicates (30 samples: 3 replicates x 10 time points) of MCF-7 cells were exposed to 10nM Estradiol for 0, 1, 2, 3, 4, 5, 6, 8, 12 , or 24 hours, and total RNA was extracted from the samples. Total RNA was used to generate paired RNA and miRNA sequencing. In analysis, We used TPM values quantified by RSEM for gene expression values and normalized expression (RPMM) values for miRNA expression values which are provided by original paper.

3.2. Comparison with other tools using expression profiles

For the analysis of A375 malignant melanoma cell data, 20 differentially expressed miRNAs (DEmiRNAs, hereafter) are reported in the original paper. Thus, we predicted the target genes for the 20 DEmiRNA using our methods and other existing tools, and compared them according to evaluation criteria as mentioned in Methods section.

Results of 9 methods were compared: mirTarGP, mirTarGP-NoGP (mirTarGP framework without gaussian process), targetExpress-TDI (time-domain intersection), targetExpress-TDU (time-domain union), MMIA-TDU (time-domain union),

KNN-MI, MINE and targets predicted by two correlation coefficient (Pearsons correlation coefficient (PCC) [14] and Spearman correlation coefficient (SCC) [15]).

GenMir++ and Magia2 were excluded from comparison. GenMir++ produced no target genes in any dataset and Magia2 produced the target genes for some miRNAs but the result did not exceed 20 DEmiRNAs. It seemed the web-site of Magia2 was outdated because it has not been updated from 2012 when the web-site was published and the developer did not provide any other executable programs. MMIA predicted target genes for a few miRNAs at some time-points. However, there are no common pairs of miRNA and target genes for all time-points. Therefore, MMIA-TDU was only included for the futher comparison analysis.

3.3. Performance comparison with other tools on A375 malignant melanoma cell data

For the analysis of A375 malignant melanoma cell data, 20 differentially expressed miRNAs (DEmiRNAs, hereafter) are reported in the original paper. Thus, we predicted the target genes for the 20 DEmiRNA using our methods and other existing tools, and

compared them according to evaluation criteria as mentioned in Methods section.

Results of 9 methods were compared: mirTime, mirTime-noGP (mirTime framework without gaussian process), targetExpress-TDI (time-domain intersection), targetExpress-TDU (time-domain union), MMIA-TDU (time-domain union), KNN-MI, MINE and targets predicted by two correlation coefficient (Pearsons correlation coefficient (PCC) [14] and Spearman correlation coefficient (SCC) [15]).

GenMir++ and Magia2 were excluded from comparison. GenMir++ produced no target genes in any dataset and Magia2 produced the target genes for some miRNAs but the result did not exceed 20 DEmiRNAs. It seemed the web-site of Magia2 was outdated because it has not been updated from 2012 when the web-site was published and the developer did not provide any other executable programs. MMIA predicted target genes for a few miRNAs at some time-points. However, there are no common pairs of miRNA and target genes for all time-points. Therefore, MMIA-TDU was only included for the futher comparison analysis.

As shown in Figure 3A, the target genes of 20 DEmiRNAs were produced for 9 tools: SCC0.5 (2,781), targetExpress-TDU (1,398), targetExpress-TDI (1,002), MINE (357), KNN-MI (250), mirTime

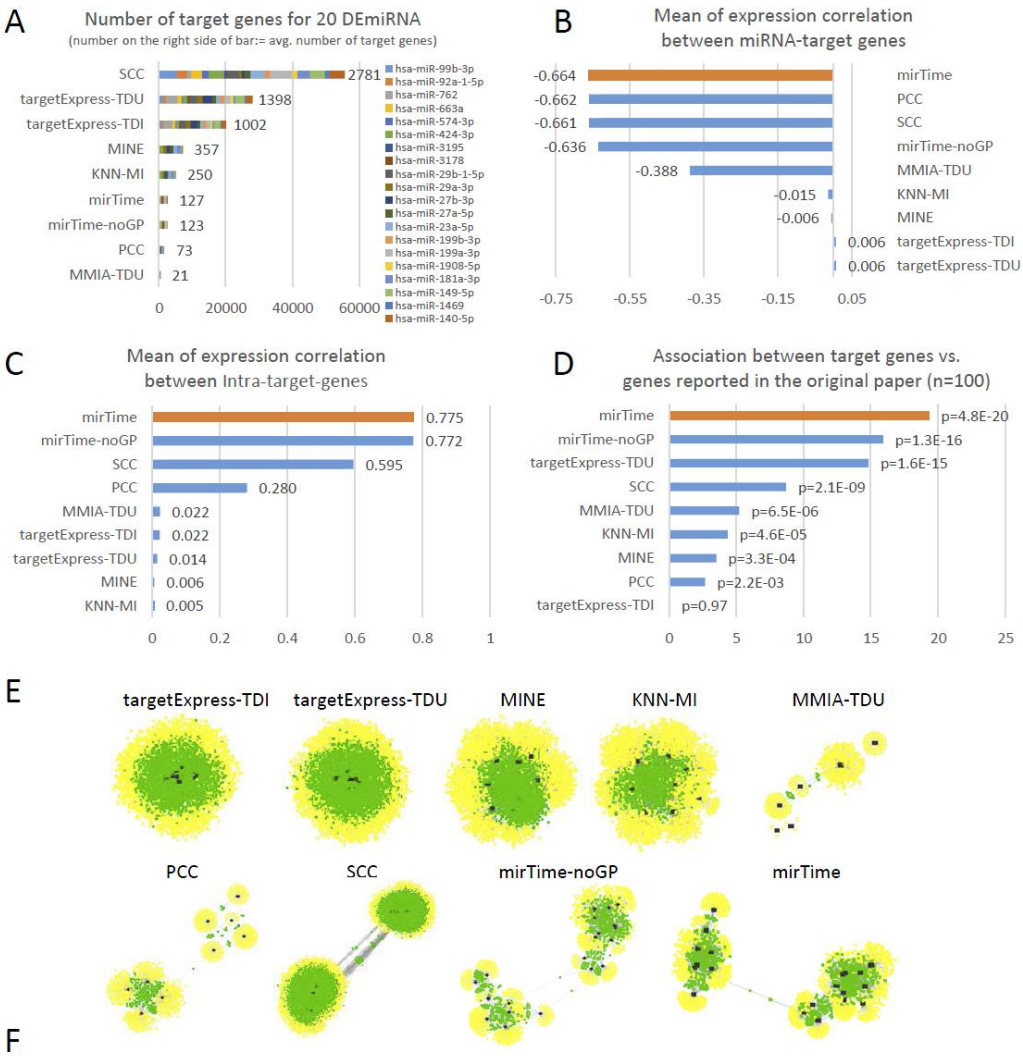
(127), mirTime-noGP (123), PCC (73), and MMIA-TDU (21), where the number in parenthesis is the average number of target genes.

3.3.1. Correlation-based evaluation

As shown in Figure 3B, the mean of the expression correlation between miRNA and target genes was computed and compared for nine methods: mirTime (-0.664), PCC (-0.662), mirTime-noGP (-0.661), SCC (-0.636), MMIA-TDU (-0.388), KNN-MI (-0.015), MINE (-0.006), targetExpress-TDU (0.006), and targetExpress-TDI (0.006), where the number in parenthesis is the mean of correlation. mirTime showed the lowest negative correlation coefficient among the nine tools. Since miRNAs suppress the expression level of the target gene, this result shows that mirTime best predicted the negative regulatory relationship for this data.

As shown in Figure 3C, the mean of the expression correlation between the target genes within the miRNA-target module was calculated and compared for nine methods: miRTime (0.775), mirTime-noGP (0.772), SCC (0.595), PCC (0.280), MMIA-TDU (0.022), targetExpress-TDI (0.022), and targetExpress-TDU (0.014), MINE (0.006), and KNN-MI (0.005), where the number in

parenthesis is the mean of correlation. mirTime showed the highest mean correlation coefficient among the nine tools.



F

Subtype	GO term	#genes	p-value
module1	endoplasmic reticulum unfolded protein response	9	5.7E-08
	positive regulation of exosomal secretion	5	3.3E-06
module2	hair follicle development	4	2.9E-04
	regulation of transforming growth factor beta2 production	2	4.4E-04
module3	positive regulation of hydrogen peroxide catabolic process	2	5.0E-05
	tricarboxylic acid cycle	4	7.5E-05
module4	positive regulation of phagocytosis, engulfment	3	1.5E-04
	negative regulation of proteolysis involved in cellular protein catabolic process	2	2.0E-04
module5	regulation of nitrogen utilization	2	1.2E-04
	post-embryonic development	6	1.4E-04
module6	frontal suture morphogenesis	3	3.1E-06
	response to organic cyclic compound	5	5.7E-05
module7	integrin-mediated signaling pathway	6	1.4E-05
	negative regulation of apoptotic process	13	1.6E-05
module8	positive regulation of osteoblast differentiation	5	1.9E-05
	negative regulation of interleukin-10 secretion	2	9.4E-05

Figure 3. Performance comparison with other existing tools for the A375 dataset. (A) The number of target genes for 20 DEmiRNAs. The number on the right side of the bar denotes average number of target genes. (B) The mean of expression correlations between miRNAs and target genes. (C) he mean of expression correlations between target genes for each miRNA. (D) Association between target genes and genes that were reported in the original paper (n = 100). (E) Visualization of miRNA−target gene networks. The black nodes are miRNA and the yellow and green nodes are the predicted target genes by each method. The green genes are co−targeted genes by at least two miRNAs. (F) The enrichged GO term for eight modules that were provived from mirTime's network

3.3.2. Target gene-based evaluation

As shown in Figure 3D, the association between the predicted target genes and those genes that was reported in origianl paper [14] was investigated and compared for nine methods: mirTime (4.8E-20), mirTime-noGP (1.3E-16), targetExpress-TDU (1.6E-15), SCC (2.1E-9), MMIA-TDU (6.5E-6), KNN-MI (4.6E-5), MINE (3.3E-4), PCC (2.2E-3), and targetExpress-TDI (0.97), where the number in parenthesis is the p-value for Fisher's exact test. mirTime showed the highest association between target genes and paper-reported genes.

3.3.3. Network-based evaluation

As shown in Figure 3E, the miRNA-target gene network was visualized for the nine methods. mirTime showed cluster structures in the network visualization. Thus, we performed network clustering and produced the five sub-networks. We investigated the biological functions of each module by performing the GO enrichment analysis as shown in Figure 3F.

3.4. Performance comparison with other tools on

MCF-7 breast cancer cell data

For the analysis of MCF-7 breast cancer cell data, 9 DEmiRNAs are reported in the original paper. Thus, we predicted the target genes for the 9 DEmiRNA using our methods and other existing tools, and compared them according to evaluation criteria as mentioned in Methods section.

Results of 9 methods were compared: mirTime, mirTime-noGP (mirTime framework without gaussian process), targetExpress-TDI (time-domain intersection), targetExpress-TDU (time-domain union), GenMir++, KNN-MI, MINE and targets predicted by two correlation coefficient (Pearsons correlation coefficient (PCC) [14] and Spearman correlation coefficient (SCC) [15]).

MMIA and Magia2 were excluded from comparison. MMIA produced no target genes corresponding to 9 DEmiRNAs and Magia2 was excluded for the same reason as in A375 melanoma cander cell data. MMIA predicted target genes for a few miRNAs at some time-points. However, there are no common pairs of miRNA and target genes for all time-points. Therefore, MMIA-TDU was only included for the futher comparison analysis.

As shown in Figure 4A, the target genes of 9 DEmiRNAs were produced for 9 tools: targetExpress-TDU (5,055), targetExpress-

TDI (994), MINE (351), mirTime-noGP (136), mirTime (126), GenMir (124), KNN-MI (55), SCC0.5 (42) and PCC (33) where the number in parenthesis is the average number of target genes.

3.4.1. Correlation-based evaluation

As shown in Figure 4B, the mean of the expression correlation between miRNA and target genes was computed and compared for nine methods: PCC (-0.615), mirTime (-0.606), mirTime-noGP (-0.589), SCC (-0.535), GenMir++ (-0.085), MINE (-0.056), targetExpress-TDI (-0.032), targetExpress-TDU (-0.017) and KNN-MI (-0.015), where the number in parenthesis is the mean of correlation. mirTime showed the lowest negative correlation coefficient following PCC.

As shown in Figure 4C, the mean of the expression correlation between the target genes within the miRNA-target module was calculated and compared for nine methods: miRTime (0.837), mirTime-noGP (0.829), SCC (0.578), PCC (0.532), KNN-MI (0.054), MINE (0.033), targetExpress-TDI (0.029), GenMir++ (0.029) and targetExpress-TDU (0.020), where the number in parenthesis is the mean of correlation. mirTime showed the highest mean correlation coefficient among the nine tools.

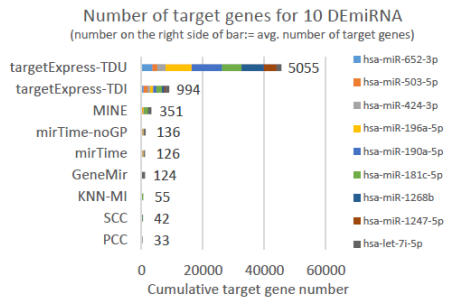
3.4.2. Target gene-based evaluation

As shown in Figure 4D, the association between the predicted target genes and those genes that was reported in origianl paper [24] was investigated and compared for nine methods: mirTime (1.9E-119), mirTime-noGP (1.6E-113), targetExpress-TDU (3.8E-96), MINE (1.4E-26), KNN-MI (7.5E-21), targetExpress-TDI (2.7E-17), GenMir++ (4.6E-11), SCC (1.5E-10) and PCC (1.3E-4), where the number in parenthesis is the p-value for Fisher's exact test. mirTime showed the highest association between target genes and paper-reported genes.

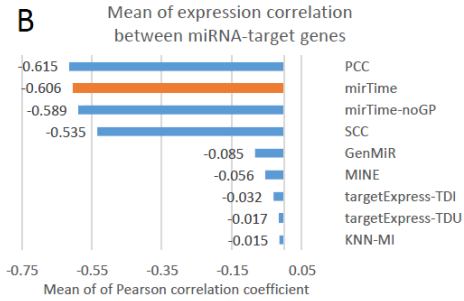
3.4.3. Network-based evaluation

As shown in Figure 4E, the miRNA-target gene network was visualized for the nine methods. mirTime showed cluster structures in the network visualization. Thus, we performed network clustering and produced the five sub-networks. We investigated the biological functions of each module by performing the GO enrichment analysis as shown in Figure 4F.

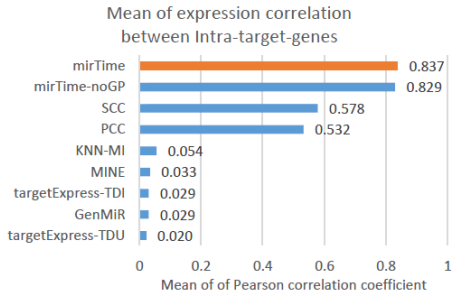
A



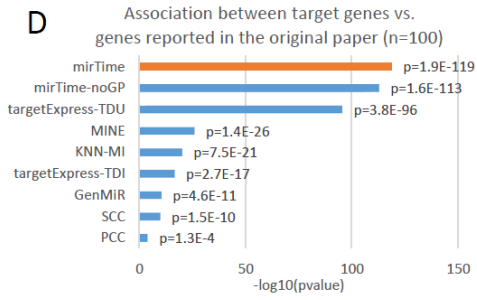
B



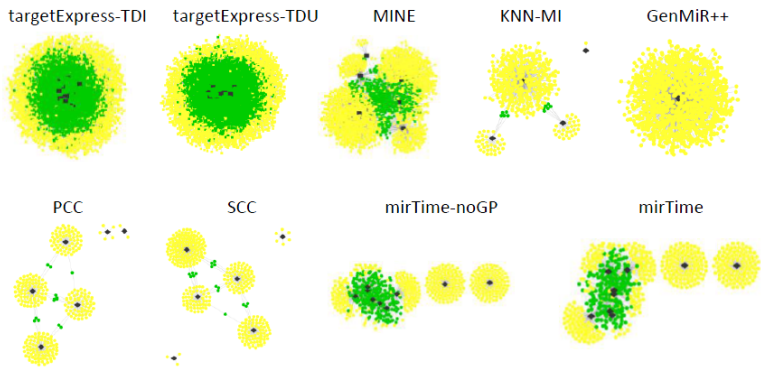
C



D



E



F

Subtype	DEmiRNAs	GO term	#genes	p-value
module1	hsa-miR-424-3p hsa-miR-652-3p hsa-miR-1247-5p	semaphorin-plexin signaling pathway involved in axon guidance	3	2.2E-05
		regulation of cell shape	6	1.2E-04
module2	hsa-miR-1268b	DNA replication	11	1.9E-10
		DNA repair	13	3.7E-10
module3	hsa-let-7i-5p	extrinsic apoptotic signaling pathway in absence of ligand	4	8.7E-06
		response to organic cyclic compound	4	4.5E-05
module4	hsa-miR-503-5p	cellular response to indole-3-methanol	2	2.3E-04
		protein import	2	2.3E-04
module5	hsa-miR-181c-5p hsa-miR-196a-5p	ventricular septum development	4	1.6E-04
		regulation of lipopolysaccharide-mediated signaling pathway	2	2.7E-04
module6	hsa-miR-190a-5p	IRE1-mediated unfolded protein response	7	1.7E-08
		protein exit from endoplasmic reticulum	3	3.5E-06

Figure 4. Performace comparison with other existing tools for the MCF7 dataset. (A) The number of target genes for 9 DEmiRNAs. The number on the right side of the bar denotes average number of target genes. (B) The mean of expression correlations between miRNAs and target genes. (C) The mean of expression correlations between target genes for each miRNA. (D) Association between target genes and genes that were reported in the original paper (n = 100). (E) Visualization of miRNA−target gene networks. The black nodes are miRNA and the yellow and green nodes are the predicted target genes by each method. The green genes are co−targeted genes by at least two miRNAs. (F) The enrichged GO term for six modules that were provived from mirTime's network

Chapter 4. Discussion

In this paper, we propose a new pipeline, mirTime, that predicts microRNA targets by integrating sequence feature and time-series expression profiles in specific experimental condition. For two datasets with different experimental conditions and cell types, condition specific target modules were predicted with our new pipeline for differentially expressed microRNAs that were reported from original papers. The context specificity of target modules was measured with three (statistics, correlation-based, target gene-based, network-based) evaluation criteria. MirTime showed good performance in three criteria than other microRNA target prediction methods using expression profiles.

For the putative targets from sequence based target prediction tools or experimentally validated databases, nearly one third or more of them have shown positive correlation in real time-series data. With our new pipeline, condition specific target genes showing

negative correlation with miRNAs can be extracted from the static pool of putative targets. Specifically, hsa-miR-503-5p, reported as estrogen-responsive miRNA, showed different target distribution in between A375 malignant melanoma cell data and MCF7 breast cancer cell data.

The performance of the algorithm has been slightly improved by using the GP-weight vector, rather than using the raw expression vector as shown in Results section. The silhouette scores of the clusters after performing spherical k-means were also increased when the GP-weight vector is used.

With the proposed algorithm in this paper, it is expected that the target genes of miRNA specific to each experimental environment are predicted. Since the result is limited to two datasets, it is necessary to use and verify this tool for more time-series data in the future. For the further research, it is considered that performance of the algorithm would be advanced considering the effect of different clustering methods and correlation measurements between genes and miRNAs besides spherical K-means and PCC.

Bibliography

1. Bartel, D.P.: MicroRNAs: target recognition and regulatory functions. Cell 136, 215−233 (2009)

2. Lewis, B.P., Burge, C.B., Bartel, D.P.: Conserved seed pairing, often anked by adenosines, indicates that housands of human genes are microRNA targets. Cell 120, 15−20 (2005)

3. Michael, K., Nicola, I., Ulrich, U., Ulrike, G., Eran, S.: The role of site accessibility in microRNA target recognition. Nature Genetics 39, 1278−1284 (2007)

4. Azra, K., Dominic, G., Matthew, N.P., Rachel, W., Lauren, R., Eric, J.E., Philip, M., Isabelle, d.P., Kristin, C.G., Markus, S., Nikolaus, R.: Combinatorial microRNA target predictions. Nature Genetics 37, 495−500 (2005)

5. Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., Rigoutsos, I.: A pattern−based method for the identi_cation of microrna binding sites and their corresponding heteroduplexes. Cell 126(6), 1203−1217 (2006)

6. Agarwal, V., Bell, G.W., Nam, J., Bartel, D.P.: Predicting e_ective

microRNA target sites in mammalian mRNAs. eLife 4(e05005) (2015)

7. Ovando-V_azquez, C., Lepe-Soltero, D., Abreu-Goodger, C.: Improving microrna target prediction with gene expression pro_les. Nucleic Acids Res. 17, 364 (2016)

8. Chae, H., Rhee, S., Nephew, K.P., Kim, S.: Biovlab-mmia-ngs: microrna-mrna integrated analysis using high-throughput sequencing data. Bioinformatics. 31(2), 265-7 (2015)

9. Bisognin, A., Sales, G., Coppe, A., Bortoluzzi, S., Romualdi, C.: Magia2: from mirna and genes expression data integrative analysis to microrna-transcription factor mixed regulatory circuits (2012 update). Nucleic Acids Res. 40, 13-21 (2012)

10. Huang, G.T., Athanassiou, C., Benos, P.V.: mirconnx: condition-speci_c mrna-microrna network integrator. Nucleic Acids Res. 39, 416-23 (2011)

11. Kraskov, A., St□gbauer, H., Grassberger, P.: Estimating mutual information. Physical review E 69(6), 066138 (2004)

12. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. science 334(6062), 1518-1524 (2011)

13. Huang, J.C., Babak, T., Corson, T.W., Chua, G., Khan, S., Gallie, B.L., Hughes, T.R., Frey, B.J., Morris, Q.D.: A pattern-based method

for the identi_cation of microrna binding sites and their corresponding heteroduplexes. Nature Methods 4(12), 1045–9 (2007)

14. Grilli, A., Sciandra, M., Terracciano, M., Picci, P., Scotlandi, K.: Integrated approaches to mirnas target definition: time–series analysis in an osteosarcoma di_erentiative model. BMC medical genomics 8(1), 34 (2015)

15. Lu, L., McCurdy, S., Huang, S., Zhu, X., Peplowska, K., Tiirikainen, M., Boisvert, W.A., Garmire, L.X.: Time series mirna–mrna integrated analysis reveals critical mirnas and targets in macrophage polarization. Scientitic reports 6, 37446 (2016)

16. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al.: Gencode: the reference human genome annotation for the encode project. Genome research 22(9), 1760–1774 (2012)

17. Reinsbach, S., Nazarov, P.V., Philippidou, D., Schmitt, M., Wienecke–Baldacchino, A., Muller, A., Vallar, L., Behrmann, I., Kreis, S.: Dynamic regulation of microrna expression following interferon–induced gene transcription. RNA biology 9(7), 978–989 (2012)

18. Nazarov, P.V., Reinsbach, S.E., Muller, A., Nicot, N., Philippidou, D., Vallar, L., Kreis, S.: Interplay of micrornas, transcription factors and target genes: linking dynamic expression changes to function. Nucleic acids research 41(5), 2817–2831 (2013)

19. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13(11), 2498–2504 (2003)

20. Blondel, V.D., Guillaume, J.–L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), 10008 (2008)

21. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the uni_cation of biology. Nature genetics 25(1), 25 (2000)

22. Jung, D., Kim, B., Freishtat, R.J., Giri, M., Ho_man, E., Seo, J.: mirtarvis: an interactive visual analysis tool for microrna–mrna expression pro_le data. In: BMC Proceedings, vol. 9, p. 2 (2015). BioMed Central

23. Benilton, S.C., Rafael, A.I.: A framework for oligonucleotide microarray preprocessing. Bioinformatics 26(19), 2363{7 (2010). doi:10.1093/bioinformatics/btq431

24. Baran–Gale, J., Purvis, J.E., P, S.: An integrative transcriptomics approach identies mir–503 as a candidate master regulator of the estrogen response in mcf–7 breast cancer cells. RNA 22, 1592~1603 (2016)

요약

마이크로RNA(microRNA, miRNA)는 Argonaute 단백질과 결합하여 RNA-핵산가수분해소 복합체 (RNA induced silencing complex, RISC)를 이룬 후 표적 전령RNA에 붙어 표적을 절단하거나 번역(translation)을 억제하는 작용을 돕는 짧은 RNA를 말한다[1]. 성숙한 마이크로RNA는 약 20-22 염기서열의 짧은 RNA이며 5' 말단의 시드 영역(seed region)을 가지고 그와 상보적인 염기서열을 갖는 전령RNA(messenger RNA, mRNA)의 3' 비해석부위(3'-untranslated region, 3' UTR)에 결합한다고 알려져 있다. 시드 영역은 약 7개의 짧은 염기서열로 이루어진 영역으로, 이 영역에 상보적인 염기서열을 갖는다는 정보만을 이용해 표적 mRNA를 예측하게 되면 긍정 오류(false positive)가 매우 커지게 된다.

따라서 이를 개선하기 위한 많은 알고리즘 개발 연구가 수행 되어 왔는데, 가장 대표적인 것으로는 TargetScan이 있다. 최근의 TargetScan 논문에서는 seed region 정보 외에도 3' 말단에 추가로 존재하는 표적 mRNA와의 상보적 결합 정보, 표적 mRNA가 갖는 3' UTR의 길이, 해당 결합의 열역학적 안정성 등의 특징을 가지고 선형 회귀 모형을 학습함으로써 표적 예측 알고리즘을 개선하였다. 그러나 이러한 염기 서열 기반 알고리즘들은 특정 세포주(cell line)나 실험 조건마다 세포 내의 전사체(transcriptome)의 발현이 달라지게 되고, 따라서 miRNA의 표적 mRNA도 달라질 수 있다는 점을 간과하고 있다는 단점이 있다. 따라

서 최근 한 단계 더 나아가 각 샘플의 발현량 정보를 추가로 이용한 표적 예측 알고리즘들도 존재한다. 가장 대표적인 것은 TargetExpress로, 여러 샘플에서의 발현 프로파일을 이용해 Support Vector Machine(SVM)을 학습함으로써 정확도를 높였다. 그러나 현재 miRNA와 mRNA의 시계열 발현량 데이터를 이용한 표적 예측 알고리즘은 상대적으로 연구가 많이 되어 있지 않다. 시계열 발현량 데이터는 miRNA와 유전자(Gene) 간 음의 상관관계 정보를 직접적으로 유추해 낼 수 있고, 타겟 예측 후 생물학적으로 유의미한 결과를 도출해 낼 수 있는 추가 분석이 용이하기 때문에 이를 이용한 표적 예측 알고리즘 연구가 필요하다고 판단된다.

이 논문에서는 먼저 염기 서열 정보를 이용하여 각 miRNA가 표적하는 유전자 군집을 뽑은 다음, 그 군집에 Spherical K-means 알고리즘과 시계열 데이터로 학습시킨 가우시안 프로세스 모델(Gaussian Process Model)를 이용해 miRNA의 표적 유전자 모듈을 찾는 과정을 수행하였다. 일반적으로 생물학 실험에서 나온 시계열 데이터들은 시간 축의 수가 적고 간격이 일정하지 않아 분석이 어렵다는 단점이 있는데, 본 논문에서는 가우시안 프로세스 회귀 모델을 이용함으로써 그러한 난점을 극복하였다.

주요어 : 가우시안 프로세스 회귀 모델, 마이크로RNA, 표적 예측, 시계열 데이터 분석

학번 : 2016-20460

감사의 글

짧고도 긴 석사 기간 동안 저에게 많은 도움을 주신 분들께 정말 감사하다는 말을 전하고 싶습니다. 먼저 부족한 저를 이끌어주시고 많은 조언과 지원을 해주신 김 선 교수님께 가장 감사드립니다. 교수님의 지도 덕분에 석사를 잘 끝마칠 수 있었으며 연구에 대해 많은 것을 배울 수 있었습니다. 또한 항상 저를 믿어주시고 의지할 곳이 되어주신 부모님께도 정말 감사드리고, 앞으로 실망시켜 드리지 않는 딸이 되도록 하겠습니다. 마지막으로 여러 방면으로 저를 도와주신 연구실 식구들 감사합니다. 연구와 대학원 생활에 있어서 많은 도움이 되어주신 겨리 선배, 홍렬 선배, 인욱 선배와 힘들 때마다 이야기를 나눠 준 애란이, 그리고 301동 동료들에게 특별히 감사의 말을 전하고 싶습니다. 그 외에도 저를 도와주신 모든 분들 덕분에 석사를 무사히 마무리하게 되었습니다. 앞으로 저도 누군가에게 도움이 될 수 있는 사람이 되도록 하겠습니다. 감사합니다.