creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Ph.D. DISSERTATION

Network and Clustering Algorithms for the Analysis of Time Series Gene Expression Data

시계열 유전자 발현 패턴 분석을 위한 네트워크 분석 및 클러스터링 기법

AUGUST 2018

DEPARTMENT OF ELECTRICAL ENGINEERING &
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Kyuri Jo

Network and Clustering Algorithms for the Analysis of Time Series Gene Expression Data

시계열 유전자 발현 패턴 분석을 위한 네트워크 분석 및 클러스터링 기법

지도교수 김 선

이 논문을 공학박사 학위논문으로 제출함

2018 년 4 월

서울대학교 대학원

전기.컴퓨터 공학부

조 겨 리

조겨리의 공학박사 학위논문을 인준함

2018 년 6 월

위 원 장	박근수	(인)
부위원장	김선	(인)
위 원	이광근	(인)
위 원	윤성로	(인)
위 원	채희준	(인)

Abstract

Expression levels of genes at the whole genome level, especially when gene expression is measured over a time period, can be useful for characterizing biological mechanisms underlying phenotypes. Dynamic gene expression information provides opportunities to understand how organisms react in specific conditions over time. Thus, time series gene expression data is growing rapidly. However, analysis of time series gene expression data is challenging since existing methods for non-time series data need to be modified to consider the time dimension.

In my doctoral study, I developed new bioinformatics methods to analyze time series data in the context of biological network or pathway propagation. My thesis consists of three studies. In the first study, a network topology-based approach for pathway enrichment analysis, TRAP, is developed for analyzing time series transcriptome data. TRAP extends the existing pathway analysis method, SPIA, for time series analysis and estimates statistical values to measure the dynamic propagation of signaling effect in the pathway graph. In experiments on a proprietary dataset for the analysis of rice upon drought stress, TRAP was able to find relevant pathways more accurately than several existing methods. In the second study, a method to detect regulators of perturbed pathways, TimeTP, is developed. TimeTP performs pathway analysis first to determine a set of perturbed sub-pathways containing genes that are connected to propagate expression changes over time by measuring cross-correlation between two vectors of expression. To detect regulators of the perturbed pathways, TimeTP extends the gene network to include upstream regulators of genes such as transcription factors. Influence maximization technique is used to evaluate

and rank the influence of regulators on the perturbed pathways. TimeTP was applied to PIK3CA knock-in dataset and found significant sub-pathways and their regulators relevant to the PIP3 signaling pathway. In the final study, a method to infer gene network using clusters for time series gene expression data is proposed. Although several clustering methods have been developed, most of the algorithms do not take the time-to-time dependency and algorithms for time series assume evenly spaced time points. However, biological experiments are often performed in unevenly spaced time intervals due to the experimental constraints on biological samples. This study aims to incorporate Gaussian process regression into the clustering process to predict unobserved values in time series data and provide more accurate clustering result. In addition, a network of clusters is generated by measuring distance (similarity) of expression patterns between clusters. As a distance measure, shape-based distance (SBD) is used to capture similarity between time-shifted patterns. The proposed method can infer gene regulatory relationship and cascading signaling effect over time. In summary, my doctoral study analyzes gene expression time series to find perturbed pathways and sub-pathways with expression propagation over time, to detect regulators of the pathways and to cluster genes of which expression profiles are represented as Gaussian processes.

Keywords: High throughput sequencing, RNA-seq, Gene expression, time series, Biological pathway, Clustering, Gaussian process

Student Number: 2013-20884

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Gene expression

Gene is a sequence of DNA which codes for a molecule that has a specific function. In many cases, the functional product of a gene is a protein. Known as the central dogma of molecular biology, information in a gene's DNA is transmitted to a protein through two major steps, transcription and translation (Figure 1.1). In transcription, the DNA sequence of a gene is copied to make an RNA molecule. Except for some cases that RNA molecule itself serves a function within the cell, the information that the RNA molecule carries is used to produce proteins. RNA that carries messages from the DNA to other areas of the cell is called messenger RNA (mRNA). In translation, the sequence of mRNA is kept by or translated to specify the amino acid sequence of a polypeptide.

Each step of the gene expression can be a potential control point for gene

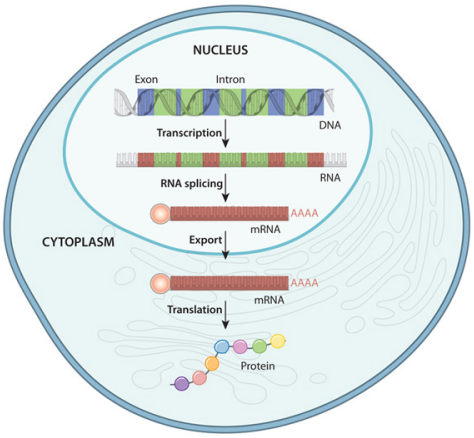


Figure 1.1: An overview of the flow of information from DNA to protein in a eukaryote [1].

regulation. Especially, the primary control point is the beginning of the protein production process, the initiation step of transcription. Transcription begins with an enzyme called RNA polymerase attaches to the DNA strand and other cell proteins called transcription factors work together with the polymerase. Transcription factors can regulate - turn on or off - genes by promoting or blocking the recruitment of polymerase to specific genes.

Gene expression in a cell can determine the cell function and even change the phenotype of an organism. Figure 1.2 is an example of the phenotypic change of tomatoes with induced expression of transcription factors. Compared to the wild type tomato in the first row, other transgenic tomatoes show a dramatic change of the phenotype and the expression levels of other genes related to metabolism and biosynthesis, due to the overexpression of transcription factors Delia and Rosea1 (the second and fourth rows) and AtMYB12 (the third and fourth rows). As shown in this example, a transcription factor can bind to and regulate a number of genes including regulators to cause the cascading impact.

Therefore, investigating the interaction among genes and how it affects the global gene expression is an important research problem.

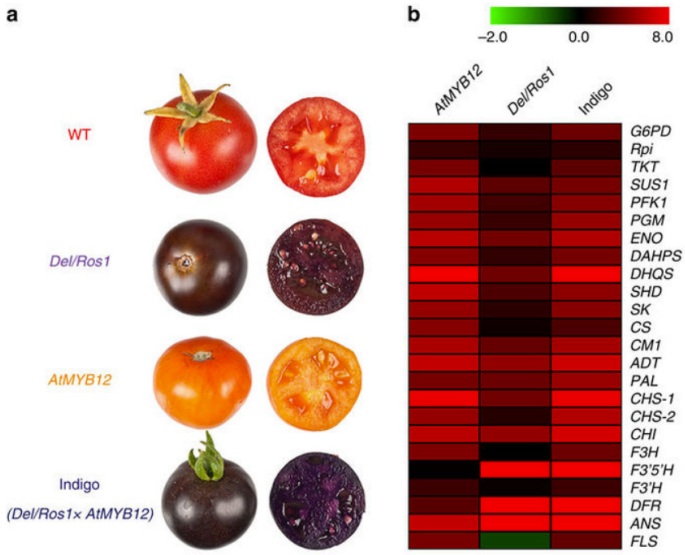


Figure 1.2: Example of the phenotypic changes by the overexpression of transcription factors (AtMYB12, Delila and Rosea1) in tomatoes [2].

1.2 Biological pathway analysis

Biological pathway is a series of molecular events in a cell that leads to a product or a change in a cellular state. Pathways can be classified into metabolic, genetic and signaling pathway. Metabolic pathway is a series of chemical reactions where a substrate binds to an enzyme and generates products that can be the next substrates. Genetic pathway indicates gene regulatory network where the regulators like transcription factors control the gene expression levels of their target mRNA and proteins. Signaling pathway is the process of signal trans-

duction that starts when a receptor senses stimuli and gives rise to a signaling cascade that is a chain of biochemical events such as protein phosphorylation. Several biological pathways are curated as databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and REACTOME.

Pathway perturbation is one of the primary research problems in systems biology because the identification of perturbed pathways can reveal dysregulated biological mechanisms that originate from stimuli or in disease conditions [3, 4, 5]. As shown in Figure 1.3, most of the pathway analysis methods use condition-specific gene expression data with two or more conditions (e.g. normal vs. cancer) to calculate the fold change of genes or detect differentially expressed genes (DEGs) first. The list of DEGs and their fold changes are integrated with the pathway information from databases and each pathway is scored and ranked by the statistical tests.

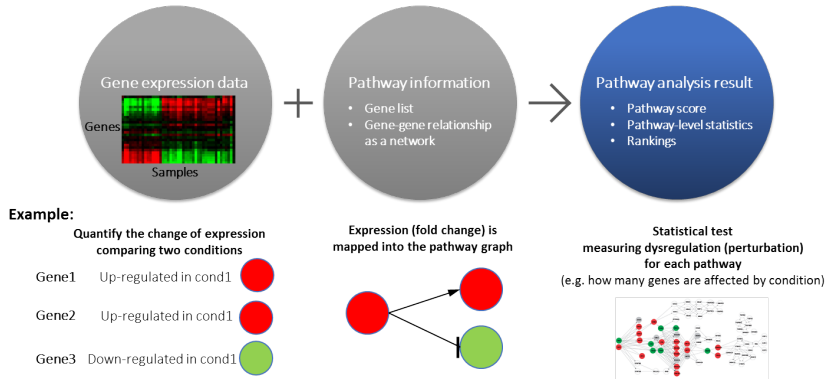


Figure 1.3: General pipeline of pathway analysis.

Timeline of pathway analysis method development is described in Figure 1.4. The early methods of pathway analysis include the gene set enrichment

analysis (GSEA) by [6] and improved versions of GSEA [7, 8] that use gene-level statistics calculated from the test of differential expression. Later, graph-based algorithms for pathway analysis were developed to utilize interaction information between genes or proteins in terms of curated pathway databases such as KEGG [9]. The graph-based pathway analysis has been developing with a seminal work called the signaling pathway impact analysis (SPIA) by [10] and the current trend is to focus on locating perturbed sub-pathways, rather than entire pathways. Tools to determine perturbed pathways include DEGraph [11], DEAP [12], and Clipper [13].

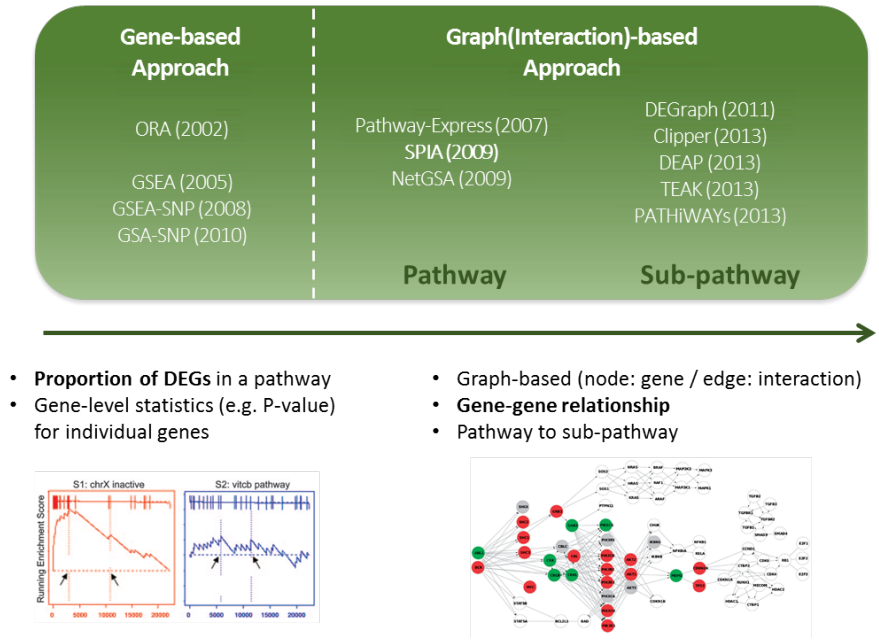


Figure 1.4: Timeline of pathway analysis method development.

1.3 Time series analysis on gene expression data

A time series is a sequence of numerical data points indexed in time order. Time series are analyzed in various fields including statistics, electrical engineering and economics. Most common form of time series used in bioinformatics is gene expression levels measured at multiple time points aiming to detect the changes triggered by the external stimuli or the onset of disease. The most important and challenging characteristic of gene expression time series is that genes interact with each other, thus a correlation between their time series exists. For example, the expression levels of transcription factor and its binding target genes can show similar patterns with or without time delay. In addition, time series of gene expression is often measured in multiple conditions or phenotype such as wild type and transgenic plants which adds an additional dimension on the time series.

Analyzing transcriptome data can be done in many different ways for different purposes. Thus, there are numerous computational methods in different categories such as (1) methods for analyzing time series transcriptome data, (2) methods for the pathway based analysis of time series data, and (3) methods for identifying regulators while analyzing time series data.

1.3.1 Methods for analyzing time series transcriptome data

Considering the time dimension, identifying perturbed (sub-)pathways in the time series transcriptome data is much more challenging. Due to the computational challenges, many computational methods for the time series analysis do not utilize pathway information directly. Widely used time series analysis methods employ a strategy of finding differentially expressed genes (DEGs) by fitting the gene expression data to a model with distributional assumption such

as Gaussian or negative binomial distribution. By utilizing statistical methods such as ANOVA [14], several tools and algorithms [15, 16, 17] have been developed for detecting DEG from time series microarray data. With the emergence of next-generation sequencing data, DEG detection algorithms utilized Gaussian process [18] or hidden Markov models [19] to identify DEGs from time series RNA-seq data. Instead of identifying DEGs, clustering approaches have been developed to determine a set of genes with a similar pattern of gene expression profile. Clustering expression data in the gene-time dimension is performed by considering correlation [20] or by model-based clustering methods [21, 22]. Recent methods such as [23] are further developed to handle data with higher dimensions such as gene-sample-time. The main limitation of DEG or clustering approaches is that a list of DEGs or genes in clusters requires further analysis in terms of curated knowledge such as KEGG pathways and the selection of significant pathways is usually determined by simple statistical methods such as Fisher's exact test. Thus, this analysis process does not consider curated knowledge such as relationships among genes, e.g. those in KEGG pathways, to determine how genes interact over time. More advanced methods consider relationship between genes or between time points (e.g. dynamic Bayesian network) to infer the gene regulatory network [24] or protein-protein interaction network [25]. These methods, although powerful, are limited to the analysis of small size gene sets [26].

1.3.2 Methods for pathway based analysis of time series data

To incorporate pathway information for the time series data analysis, pathways are modeled as graphs. Two recent graph-based pathway analysis algorithms for time series data are TRAP [27] proposed in this thesis and TimeClip [28]. TRAP extends the existing pathway analysis method, SPIA [10], for time series analysis

and estimates statistical values to measure the dynamic propagation of signaling effect in the pathway graph. Detail of TRAP will be described in Chapter 2. TimeClip [28] employs a junction tree algorithm to form sub-pathways using the same method used in Clipper [13] to determine significant sub-pathways in terms of the first principal component from the gene expression data. Although these algorithms produces a list of biological processes with significant changes over time, few attempts have been made to locate the regulator that initiates the pathway perturbation.

1.3.3 Methods for identifying regulators while analyzing time series data:

DREM [29] is an example of incorporating regulators in clustering analysis. It estimates transcription factors (TFs) regulating a cluster by Input-Output Hidden Markov Model, but it is hard to discover biological implication from the result due to the clusters with multiple or overlapping biological functions. Master regulator analysis (MRA) [30] introduces a method to rank TFs in the gene regulatory network, but not considering dynamic expression profiles of genes.

1.4 Computational challenges and solutions in time series gene expression

Although several algorithms and tools have been developed for time series gene expression data, there are still remaining challenges. My doctoral study includes three different methods to solve those problems step by step from pathway analysis to sub-pathway analysis and from using discrete time points to considering continuous time points and their dependencies (Figure 1.5). Detailed description of the problems and solutions are as follows.

- **A network topology-based approach for pathway enrichment analysis of time series transcriptome data (TRAP) [27]** : Most of the existing algorithms for pathway analysis do not consider time dimension of the gene expression data. TRAP leverages the technique similar to SPIA [10] to detect and additionally quantify pathways with a significant expression propagation along the pathway graph in the time order.

- **Detecting regulators in a network labeled with time series by cross-correlation and influence maximization technique (TimeTP) [31]** : TimeTP narrows the focus to sub-pathways with genes connected to propagate expression changes over time by measuring cross-correlation between two vectors of expression. TimeTP leverages influence maximization technique to detect and evaluate regulators of the perturbed pathways. In addition, TimeTP generates TF-Pathway map in time clock that enables user to navigate the perturbation propagation route along time.

- **Inference of cluster network from unevenly spaced time series by Gaussian process and shape-based clustering (TiClNet)** : TiClNet introduces Gaussian process regression to time series clustering to consider unobserved time points and their dependencies, improving the accuracy of clustering. After clustering, TiClNet generates a network of clusters by overlapping expression patterns and biological functions between clusters, which can be more informative than typical clustering results for biological research.

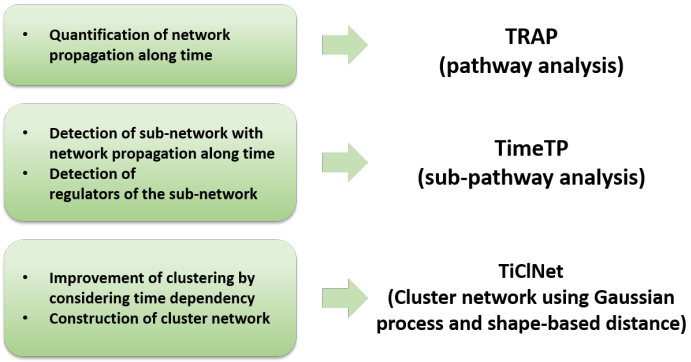


Figure 1.5: Computational challenges and solutions in time series gene expression.

1.5 Outline of the thesis

Chapters 2, 3 and 4 introduce independent studies related to network and clustering algorithms for the analysis of time series gene expression data. In Chapter 2, a network topology-based approach for pathway enrichment analysis, TRAP, is described, which aims to extend the existing method for time series and quantify network propagation along time. Chapter 3 describes TimeTP, a method to detect sub-network with network propagation along time and regulators of the sub-network. Chapter 4 suggests a method to improve clustering by time point estimation using Gaussian process regression and construct the cluster network.

Chapter 5 summarizes the studies with the results so far and expected results of the ongoing study. The thesis is concluded by an appendix the bibliography of the cited references.

Chapter 2

A network topology-based approach for pathway enrichment analysis of time series transcriptome data

Measuring expression levels of genes at the whole genome level can be useful for many purposes, especially for revealing biological pathways underlying specific phenotype conditions. When gene expression is measured over a time period, I have opportunities to understand how organisms react to stress conditions over time. Thus many biologists routinely measure whole genome level gene expressions at multiple time points. However, there are several technical difficulties for analyzing such whole genome expression data. In addition, these days gene expression data is often measured by using RNA-sequencing rather than microarray technologies and then analysis of expression data is much more complicated since the analysis process should start with mapping short reads and produce differentially activated pathways and also possibly interactions

among pathways. In addition, many useful tools for analyzing microarray gene expression data are not applicable for the RNA-seq data. Thus a comprehensive package for analyzing time series transcriptome data is much needed. In this article, I present a comprehensive package, time series RNA-seq Analysis Package (TRAP), integrating all necessary tasks such as mapping short reads, measuring gene expression levels, finding differentially expressed genes (DEGs), clustering and pathway analysis for time series data in a single environment. In addition to implementing useful algorithms that are not available for RNA-seq data, I extended existing pathway analysis methods, ORA and SPIA, for time series analysis and estimates statistical values for combined dataset by an advanced metric. TRAP also produces visual summary of pathway interactions. Gene expression change labeling, a practical clustering method used in TRAP, enables more accurate interpretation of the data when combined with pathway analysis. I applied my methods on a real dataset for the analysis of rice (Oryza sativa L. Japonica nipponbare) upon drought stress. The result showed that TRAP was able to detect pathways more accurately than several existing methods. TRAP is available at http://biohealth.snu.ac.kr/software/TRAP/.

2.1 Background

Biological processes involve a set of genes that interact dynamically and differentially in specific conditions. Some of the genes are activated and suppressed in condition specific manners and gene activation/suppression affects regulation of other genes directly or indirectly. Thus understanding biological processes from transcriptome (whole genome gene expression) data is a very important analysis task. One of the most effective methods to understand biological processes is to match up genes that are activated and suppressed in terms of curated

pathway databases such as KEGG [9] or DAVID [32]. Techniques for measuring transcriptome have rapidly changed from microarray to sequencing as the cost for sequencing cost decreases. The high throughput RNA sequencing (RNA-seq) technique has advantages over microarray technique such as the decreased noise level and more replicable results [33], compared to microarrays. For this reason, the number of time series RNA-seq data sets in the public domain has increased dramatically over the past few years [34]. The main issue with RNA-seq is to handle sequencing data that is much bigger than microarray data. When transcriptome data is measured over time, analysis of RNA-seq data requires to consider the time dimension. Thus, analysis of whole genome transcriptome data is a quite involved task. Two major hurdles are:

- A number of tools developed for microarray data are not suitable for the RNA-seq data.
- Analysis of RNA-seq data requires use of multiple tools in a pipeline. Especially analysis of time series RNA-seq data is very complicated, thus only bioinformatics experts can construct such pipeline.

To address these challenges, I developed a comprehensive package, time series RNA-seq Analysis Package (TRAP), for analyzing time series transcritptome data. There have been many packages developed for analyzing time series gene expression data [33]. The most widely used technique is to identify DEGs. Packages for finding DEGs from time series data include SAM [35], LIMMA [36], EDGE [37], maSigPro [17] and BETR [38]. However, most of them are developed for microarray data and they assume that gene expression follows normal distribution. Users with RNA-seq data, therefore, should perform additional conversion process or use the certain type of distribution (e.g., Poisson or Binomial) which needs further validation. Some of the package does not support time

series analysis for RNA-seq data, some have constraints on the number of replicates and some are out dated, not compatible with current operating systems. Another important analysis topic is clustering genes that have similar time series expression patterns. Several clustering packages are available. CAGED [21] uses autoregressive equations, GQL [22] uses Hidden Markov Model, STEM [39] uses expression profiles, TimeClust [40] uses stochastic models, DynaMiteC [41] uses impulse models, and PESTS [42] uses various features selected by user.

Finding DEGs and clustering is the first step for identifying genes that may have an important role in relation to phenotypes. However, the analysis needs to go at least one step further to extract biological implication from the gene list. The most widely used method for this additional analysis step is pathway analysis. Through research on biological process over several decades, I now have significantly accumulated knowledge about the biological pathways, genes and proteins, and their interactions. The pathway knowledge is available as public repositories such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [9] or Gene Ontology [43]. By combining the pathway information and a list of significant genes from the experimental results, inference of active pathways is possible and this method is called knowledge base-driven pathway analysis [3]. Most of the packages for DEGs and clustering introduced above do not include pathway analysis functions. Instead, there have been studies of inferring pathways from time series expression data using prior knowledge and DBN [44], Granger causality and Minimum Spanning Tree [45], Ordinary Differential Equation [46]. Albeit these studies have given us important algorithms and findings for pathway inference, they are not of practical use, as the source code or software for their analysis are not available and the analysis result do not include pathways from public databases.

On the other hand, several knowledge base driven pathway analysis algo-

rithms using public databases have been developed over time and they include Over-representation Analysis (ORA) and Pathway Topology (PT)-based analysis [3]. ORA [47] evaluates pathways by using DEGs and ORA uses GO database to show biological process underlying the phenotype. However, the biological process from ORA does not use specific biological pathway information such as KEGG. In addition, most of the packages are web-based or Windows-based applications which are not suitable for big data sets such as RNA-seq data. The PT-based approach utilizes the pathway network and interactions in addition to the gene expression values. SPIA [10], one of the PT algorithms, is originally developed as a component of the BioConductor package for human dataset but additional steps are needed to be used for other species.

This article describes a time series RNA-seq Analysis Package (TRAP), integrating all necessary tasks such as finding DEGs, clustering and pathway analysis for time series data. TRAP is an easy-to-use and comprehensive package since it automatically performs a series of analysis steps from mapping short reads to pathway analysis. The key features of TRAP are as follows:

- TRAP is a comprehensive package that puts together the state of the art technologies for time series gene expression data analysis such as ORA and SPIA for pathway analysis and gene expression change labeling for clustering.

- TRAP is designed for RNA-seq data and it performs all necessary analysis steps automatically from mapping RNA-seq short reads to the reference to pathway analysis to visualization of pathway interactions.

- TRAP extends pathway analysis methods, ORA and SPIA, to time series analysis and estimates statistical values for combined dataset by an advanced metric.

- TRAP is successfully used to analyze time series RNA-seq data for investigating biological mechanisms for the drought resistant rice.

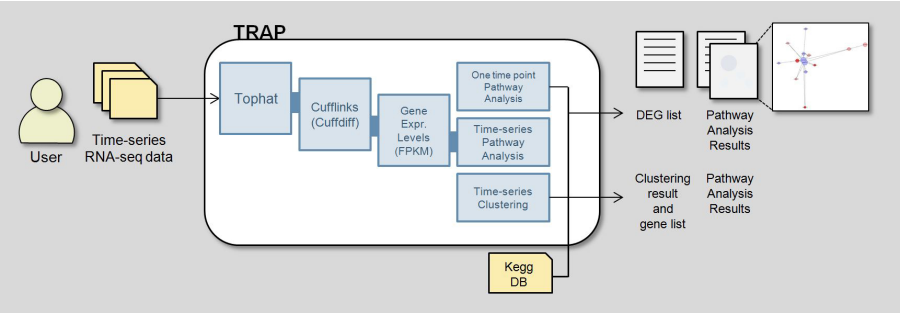


Figure 2.1: The overview of TRAP. TRAP reads the user configuration file and the location of the sequencing data. For each sample, TRAP maps paired reads to the reference genome by Tophat and estimates gene expression values by Cufflinks according to the gene annotation file. TRAP combines the gene expression values with KEGG pathway information to perform one time point pathway analysis, time series pathway analysis and time series clustering. The output of the pathway analysis includes text files with a list of DEGs and a list of pathways and image file with a graph-representation of the result. Clustering result output is a list of clusters and a list of pathways for significant clusters.

Figure 2.1 illustrates the analysis procedure of TRAP. TRAP reads the user configuration file that includes a list of sample names, location of the sequenced reads, and user-defined thresholds for selecting DEGs. After preprocessing, a series of analysis steps of TRAP as described in a white box of Figure 2.1 are executed. For each sample, paired reads are mapped to the reference genome by Tophat [48]. Using the mapping results, TRAP estimates gene expression levels by Cufflinks [49] according to the gene annotation file from Rice An-

Table 2.1: Resources used in TRAP. Tophat and Cufflinks are used for mapping reads to the reference genome and estimating gene expression values, respectively. Cuffdiff is optional for users who is to find DEGs not by cutoff. NetworkX package is imported to Python for graph visualization of the result. The rice genome annotation from RAP-DB was used to set location of the genes using Cufflinks. For the pathway analysis, ORA and SPIA methods are used and modified for time series data.

Type	Name	Usage	Ref.
Tool	Tophat	Mapping sequenced reads to the reference genome	[48]
Tool	Cufflinks	Estimation of expression values	[49]
Tool	Cuffdiff	Finding differently expressed genes	[50]
Tool	NetworkX	Visualizing the network graph	[51]
Database	Rice Annotation Project Database (RAP-DB)	Annotation of the rice genome sequence	[52]
Database	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Public repository of pathway information	[9]
Algorithm	Over-representation Analysis (ORA)	Algorithm for pathway analysis	[47]
Algorithm	Signaling Pathway Impact Analysis (SPIA)	Algorithm for pathway analysis	[10]
Algorithm	Short time series Expression Miner (STEM)	Algorithm for clustering analysis	[39]

notation Project Database (RAP-DB) [52]. As an output of Cufflinks, gene expression levels measured in FPKM (Fragments Per Kilobase of exon per Million fragments mapped) unit are stored in the text files. TRAP combines the gene expression values with KEGG pathway information to perform three types of analysis: one time point pathway analysis, time series pathway analysis and time series clustering. The pathway information in KEGG database is described in the form of KEGG Markup Language (KGML) file and the user can define a new pathway making his own KGML file. The output of the pathway analysis is two text files and an image file. The text files contain a list of DEGs and a list of pathways with their P-values and other statistics. The image file is a graph-representation of the result for easy interpretation. Clustering result includes two text files, a list of genes and a list of significant pathways for each

cluster. Tools, database, and algorithms used in TRAP are listed in Table 2.1.

The following is the TRAP input and output structure.

1. Input

Before staring TRAP, user must fill in `conFiguretxt` file to help TRAP read the following information. If user already has Tophat or Cufflinks results, the optional fields can be ignored.

`Sample pair name`

Sample names in control-treated order. Each line corresponds to one time point.

`Sample name and file path`

Sample names and their sequencing data file path in the server. If it is the paired-end sequencing data, place the white space between the path.

`Tophat path (optional for Tophat)`

Tophat path on the server.

`Reference genome path (optional for Tophat)`

Reference genome path on the server.

`Cufflinks path (optional for Cufflinks and Cuffdiff)`

Cufflinks path on the server.

`Annotation file path (optional for Cufflinks)`

Provided by TRAP (for *Oryza sativa* L. ssp.*Japonica*). The file should be in a readable format for Cufflinks, e.g., gtf/gff.

`Gene name conversion file path`

Provided by TRAP (for *Oryza sativa* L. ssp.*Japonica*). If the user-defined annotation file is used and the gene names in the annotation

file is different from those in KEGG database, the conversion file is needed.

`DEG and clustering cutoff`

2. Output

The output files are stored in `TRAP_result` folder.

Pathway analysis results

`One_time_genes.txt` : A list of DEGs from each time point.

`One_time_pathways.txt` : A list of pathways and their size, P-values, status from one time point analysis.

`time series_genes.txt` : A list of DEGs from time series analysis.

`time series_pathways.txt` : A list of pathways and their size, P-values, status from time series analysis.

Clustering results

`Clustering_genes.txt` : A list of clusters and genes in the clusters.

`Clustering_pathways.txt` : Pathway analysis result for each cluster.

2.2 Methods

2.2.1 One time point pathway analysis

The first analysis in TRAP is finding significant pathways for each time point. TRAP takes gene expression values of a certain time point as an input and executes two analysis methods to the values: ORA and SPIA. As a result, it

provides a list of DEGs and a list of pathways with the P-values from two methods. This is repeated for Lt time points independently.

The analysis starts with selecting DEGs from the genes. I define DEGs as genes which have significantly different gene expression value in the control and treated samples. For gene expression value $X(g)$ and $Y(g)$ of gene g in two different samples X and Y, I define the log fold change of FPKM values as $\Delta E(g) = log(Y(g)/X(g))$. DEGs are chosen by Cuffdiff that tests the log fold change against the null hypothesis or by picking up genes above the cutoff value defined by the user.

TRAP implemented ORA and SPIA methods for pathway analysis. ORA uses hypergeometric test comparing the ratio of total DEGs (n) to total genes (N) and the ratio of DEGs in the pathway (k) to total genes in the pathway (K) (Equation 2.1). P-value of ORA, P_{NDE}, is calculated as one minus cumulative distribution function of Equation 2.1. ORA is a simple but fundamental metric for finding significant pathways with a number of DEGs.

$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2.1}$$

SPIA uses the network topology of pathways. If a gene affects another gene, the relationship can be represented as two nodes and a directed edge between them. In this way, a pathway can be a directed graph. SPIA uses Perturbation Factor (PF) of a gene, a sum of the fold change and PF of its upstream genes g_u divided by their number of downstream genes NDS (Equation 2.2). β is 1 or -1 depending on the type of interaction, i.e., induction or repression. From PF, the fold change is subtracted to reflect only the cascading effect (Equation 2.3). Sum of *Acc(g)* of all genes in a pathway, t_A, is calculated and by using a bootstrap approach P_{PERT} is estimated (Equation 2.4-2.5). To reflect the number of DEGs to the P-value, the final P-value P_G combines P_{NDE} from

ORA and P_{PERT} from SPIA using Equation 2.6, where $c = P_{NDE} * P_{PERT}$.

$$PF(g) = \Delta E(g) + \sum_{g_u \in US} \beta \frac{PF(g_u)}{N_{DS}(g_u)} \tag{2.2}$$

$$Acc(g) = PF(g) - \Delta E(g) \tag{2.3}$$

$$t_A = \sum_{g \in P} Acc(g) \tag{2.4}$$

$$P_{PERT} = P(T_A \geq t_A | H_0) \tag{2.5}$$

$$P_G = c - c \log c \tag{2.6}$$

2.2.2 time series pathway analysis

The time series pathway analysis is the most challenging part of the TRAP analysis procedure. Although several pathway analysis methods have been developed so far, there is no method to combine data across several time points and extract one representative statistical value for each pathway. Given a set of gene expression values from Lt time points, TRAP extends two pathway analysis algorithms, time series ORA and time series SPIA, for the analysis of time series RNA-seq data. The output is a list of DEGs with annotations and a list of pathways with P-values from two analysis methods.

The analysis procedure starts in a similar way to the one time point pathway analysis in Section 2.2.1. Here I define the log fold change of FPKM values as $\Delta E_t(g) = log(Y_t(g)/X_t(g))$, where t is a certain time point. Calculation of P_{NDE} through ORA is again calculated from the cumulative distribution function of Equation 2.1, but the number of genes and DEGs are summed up from all the time points.

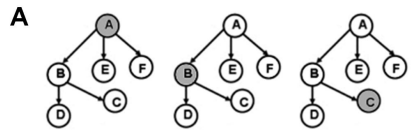
The major difference of original SPIA and time series SPIA is in the formula of Perturbation Factor of SPIA. The original SPIA assumes that downstream

genes are affected by the upstream genes at the same time point. In time series SPIA, however, upstream genes at the previous time point also affect downstream genes at the next time point, adding the additional term $\frac{PF_{t-1}(g_u)}{N_{DS}(g_u)}$ as in Equation 2.7. To adjust the ratio of effect from the previous and current upstream genes, **a time-lag factor** α and $(1-\alpha)$ is multiplied to each term. $PF(g)$ and $\Delta E(g)$ is defined as the summation of $PF_t(g)$ and $\Delta E_t(g)$ from all the time points in Equation 8. The final P-value P_G can be derived with Equation 2.3-2.6. For example, in Figure 2.2 (A) there are three time points and three genes having log fold change greater than zero. Here I assume the time-lag factor α is 1, which means the PF of downstream genes are affected only by the previous time points. At time point 1, the last sigma term of $PF_1(A)$ in Equation 2.7 will be zero because it is the first time point and there is no gene upstream of A. Therefore, $PF_1(A)$ equals $\Delta E_1(A)$ resulting in $Acc(A) = PF_1(A) - \Delta E_1(A) = 0$. At time point 2, $PF_1(A)$ is equally distributed to gene B, E and F, the downstream genes of A, resulting in $Acc(B) = Acc(E) = Acc(F) = 1$. $PF_2(B)$ becomes 3 because $\Delta E_2(B) = 2$ is added to 1. Finally at time point 3, $PF_2(B)$ is distributed to C and D as 1.5. $t_A = \sum_{g \in P} Acc(g)$ is therefore 0+1+1+1+1.5+1.5=6.

$$PF_t(g) = \Delta E_t(g) + \sum_{g_u \in US} \beta(\alpha \frac{PF_{t-1}(g_u)}{N_{DS}(g_u)} + (1-\alpha)\frac{PF_t(g_u)}{N_{DS}(g_u)}) \tag{2.7}$$

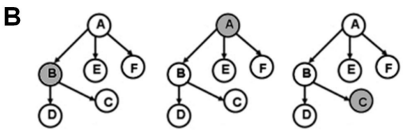
$$PF(g) = \sum_{t=1}^{L_t} PF_t(g),\ \Delta E(g) = \sum_{t=1}^{L_t} \Delta E_t(g) \tag{2.8}$$

SPIA that was originally developed to locate the signaling effect in a fixed time point, can now measure the effect through time with a change in Equation 2.7. The example in Figure 2.2 describes how time series SPIA catches the



	Time Point 1			Time Point 2			Time Point 3		
	ΔE	PF	Acc	ΔE	PF	Acc	ΔE	PF	Acc
A	3	3	0	0	0	0	0	0	0
B	0	0	0	2	3	1	0	0	0
C	0	0	0	0	0	0	2	3.5	1.5
D	0	0	0	0	0	0	0	1.5	1.5
E	0	0	0	0	1	1	0	0	0
F	0	0	0	0	1	1	0	0	0

Total Acc = 6



	Time Point 1			Time Point 2			Time Point 3		
	ΔE	PF	Acc	ΔE	PF	Acc	ΔE	PF	Acc
A	0	0	0	3	3	0	0	0	0
B	2	2	0	0	0	0	0	1	1
C	0	0	0	0	1	1	2	2	0
D	0	0	0	0	1	1	0	0	0
E	0	0	0	0	0	0	0	1	1
F	0	0	0	0	0	0	0	1	1

Total Acc = 5

Figure 2.2: A time series SPIA example with α=1. Pathway in (A) has the up-regulated genes A, B and C in time point 1, 2 and 3, respectively. There is a cascading interaction from top to bottom in accordance with time flow, which means the pathway has a stronger signaling effect. Pathway in (B) has the up-regulated genes in an inverted order of B, A and C. It has a weaker signaling effect because genes have no direct impact to downstream genes in the next time point. In effect, the sum of Acc(g) is higher in (A) than in (B), resulting in a lower P-value in A. The original figure is from [10].

flow of interactions. Figure 2.2 (A) shows the up-regulated genes A, B and C in time point 1, 2 and 3, respectively. Three genes are connected in the graph with interaction (here assumed as induction) from top to bottom of the network in accordance with time flow. It indicates that the pathway in Figure 2.2 (A) has a strong signaling effect. Up-regulated genes in Figure 2.2 (B) have an inverted order of B, A and C, which does not show the cascading effect in the pathway because genes have no direct impact to downstream genes in the next time point. As a result, the sum of $Acc(g)$ is higher in A than in B, resulting in a lower P-value in A.

2.2.3 time series clustering

time series clustering is to find groups of genes having similar expression pattern. Given a set of gene expression values from Lt time points, TRAP put the label for each gene by its expression pattern, and cluster the genes with the same label vector. This procedure is named as gene expression change labeling. Consider two gene expression values of a gene g, $X_t(g)$ and $Y_t(g)$ at the same time point. If g is up-regulated in the Y sample and the log fold change of the gene expression $\Delta E_t(g) = log(Y_t(g)/X_t(g))$ is above the threshold, then this is denoted as U. D and C are also the label which indicate down-regulated and constant, respectively. While classical clustering methods such as K-means clustering group genes with similar expression values, gene expression change labeling makes it easier to group genes by their expression pattern through time, which is more practical to identify up-regulated or down-regulated genes in terms of time points. After the formation of clusters, the significance of each cluster is calculated by the permutation based test introduced in STEM paper [39]. For each cluster, I apply the pathway analysis using ORA method to identify the biological function of the cluster. The result of time series clustering is a list of clusters and their P-values, a list of genes in the clusters, and pathway analysis result for each cluster.

The overall procedure of gene expression change labeling is as follows. For gene g in samples $X_1 \dots X_t \dots X_{L_t}$ and $Y_1 \dots Y_t \dots Y_{L_t}$, I derive the log fold change of FPKM values $\Delta E_1(g) \dots \Delta E_t(g) \dots \Delta E_{L_t}(g)$. Each gene is assigned a label for each time point t as U (Up-regulated in case sample, $\Delta E_t(g) > 0$ and $|\Delta E_t(g)|$ is above threshold), D (Down-regulated in case sample, $\Delta E_t(g) < 0$ and $|\Delta E_t(g)|$ is above threshold), or C (Constant, $|\Delta E_t(g)|$ is below threshold). The final label is the vector of the labels from L_t time points. Clusters are

formed by gathering genes which have the same label of length L_t. Pathway analysis for each cluster was performed with hypergeometric test by comparing the ratio of pathway genes (n) to total genes (N) and the ratio of pathway genes in the cluster (k) to total genes in the cluster (K).

2.2.4 Text and graph representation of pathway analysis result

The pathway analysis result in TRAP is created in both text and image forms. The text result includes a list of pathways, the size of the pathway, the number of DEGs, and their P-values (P_{NDE}, P_{PERT}, P_G) and FDR corrected P-values. The image file has an undirected graph where nodes are pathways, node size represents the number of genes in a pathway, and node color indicates P-value of a pathway. The node color is red (activated) or blue (inhibited) or green (status unknown) depending on the status of the pathway. The intensity of the node color is inversely proportional to the P-value of the pathway. Edges between nodes exist if there are common genes between pathways. Nodes are closer to each other with a larger number of common genes, which can make pathways having similar function form a cluster in the graph. This representation of pathway network in TRAP is expected to be a help to the biological research in that it enables users to examine the overall status of the pathway clusters. An example of the graph representation of the result is in Figure 2.3.

2.3 Results and Discussion

To evaluate how good TRAP is, I used rice (*Oryza sativa L. Japonica nipponbare*) mRNA-seq data generated by Illumina sequencing. The dataset compares drought-resistant AP2/EREBP transgenic rice samples with normal nontrans-

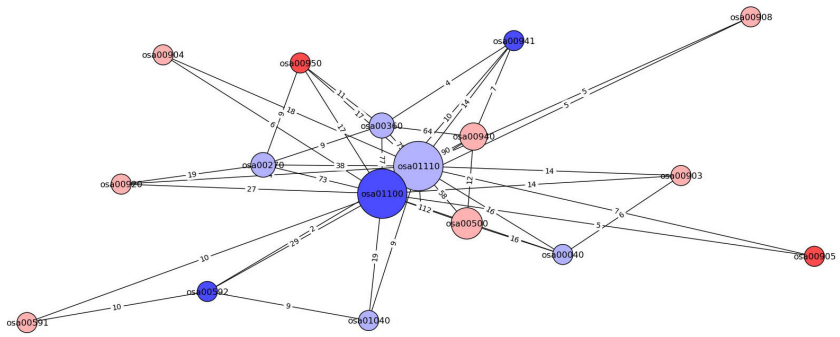


Figure 2.3: An example graph of pathway analysis result. Nodes are pathways and node size and color indicate its size and P-value. The node color is red (activated) or blue (inhibited) or green (status unknown) and the intensity of the node color is inversely proportional to the P-value of the pathway. If there are common genes between pathways, the edge is created between nodes. Nodes are closer to each other with a larger number of common genes, which can make pathways having similar function form a cluster in the graph.

genic rice samples sequenced at three time points, 0, 1, and 6 hours, after applying drought stress. Two AP2/EREBP transcription factors are amplified in rice by using overexpression vectors OsCc1:AP37 and OsCc1:AP59, respectively, resulting in increased drought resistance [53]. Drought-resistant transgenic rice samples are referred to as AP2 and normal samples as Nip.

The main use of TRAP is for the analysis of biological pathways differentially activated at each time point and for the entire experimental time period (i.e., time series). I evaluated the performance of TRAP in terms of the existing biological knowledge and also in comparison with other methods in the context of the same existing biological knowledge [10].

2.3.1 Pathway analysis results

I used TRAP to perform pathway analysis using data from the rice samples at each of 0, 1, and 6 hours and also for the whole time point vector. Table 2.2 summarizes the pathway analysis result at the 5% cutoff of the FDR adjusted P-value. If genes in a pathway interact with each other, i.e., activation or inhibition by other gene, SPIA and time series SPIA calculate P_{PERT}, P_G and P_{G-FDR} and detect differentially activated pathways by considering gene interactions in the network topology. If all genes in a pathway are independent (i.e., No gene activates or inhibits another gene), the pathway only has P_{NDE} from ORA.

Plant hormone signal transduction and *Diterpenoid biosynthesis* pathways are found to be activated both in one time point analysis and the whole time series analysis. In particular, *Plant hormone signal transduction* is significantly activated both in terms of $P_{NDE-FDR}$ and P_{G-FDR}, which indicates the pathway activation is significant not only in terms of the number of DEGs but also the cascading interaction in the pathway network. Remarkable change of gene expression is shown in abscisic acid (ABA) signaling path (Figure 2.4). ABA is a messenger in the regulation of the plant's water status that causes induction of stomatal closure in the guard cells [54]. Consequential stomatal closure reduces leaf transpiration and prevents excessive water loss in the tissues [55]. *Diterpenoid biosynthesis* is comprised of various diterpenoid generation paths such as gibberellin (GA), levopimaric acid, neoabietic acid, momilactone A, aphidicolin, etc. In particular, the path for producing GA is down-regulated in process of time while the other paths are rather up-regulated (Figure 2.5). Activation of ABA and inhibition of GA from these two pathways reflect the renowned antagonism between them [56]. In addition, an experimental study of AP2/EREBP transcription factor shows that OsAP2-39 overexpressed trans-

genic rice induced an increase of the endogenous ABA level and deactivation of GAs [57], which has parallels in my results.

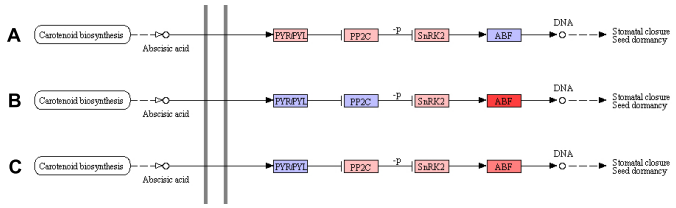


Figure 2.4: Plant hormone signal transduction pathway colored from rice data set. Path related to ABA signalling is cropped from the original pathway. A, B and C indicates 0, 1, 6 hour after drought stress. ABA signalling path is activated (colored in red) in course of time.

Biosynthesis of unsaturated fatty acids pathway appears in the result of time series pathway analysis but not at the individual time point, thus it is a potentially a good example to show the effectiveness of time series pathway analysis. The pathway is mostly up-regulated in 1 hour in drought-resistant rice (Figure 2.6), similar to the *Fatty acid biosynthesis* pathway not shown in the result but ranked as 10th important pathway among 120 pathways. Simultaneously, *alpha-Linolenic acid metabolism* pathway shows repressed beta-oxidation in course of time which leads to the reduced amount of fatty acid degradation (Figure 2.7). Previous comparative studies of cowpea and coconut tree showed that under water stress condition, the total lipid content decreased in sensitive plants while it less decreased or even increased in early stage in resistant cultivars. The main reason for the increased lipid content lies on the increased

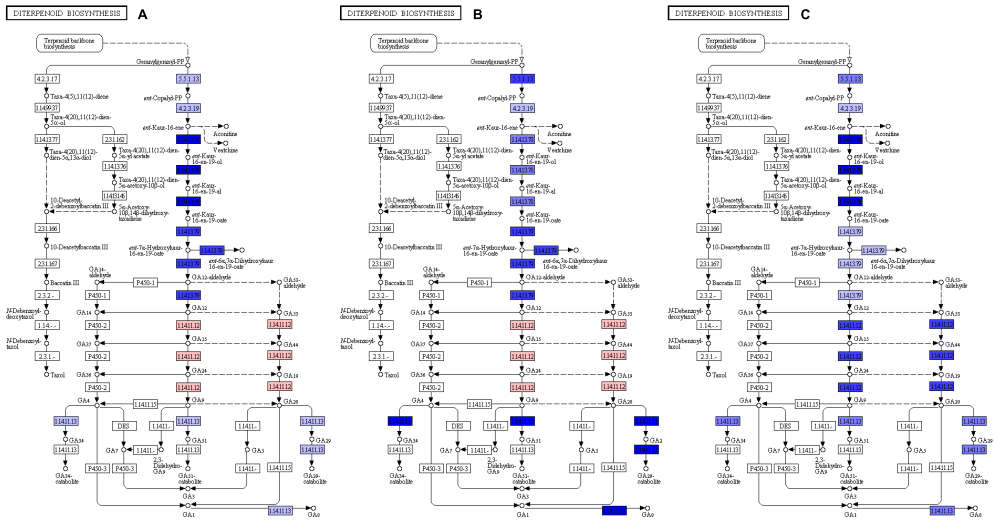


Figure 2.5: Diterpenoid Biosynthesis pathway colored from rice data set. Path related to production of gibberellin is cropped from the original pathway. A, B and C indicates 0, 1, 6 hour after drought stress. Gibberellin production path is repressed (colored in blue) in course of time.

amount of unsaturated fatty acid, especially alpha-linolenic acid (18:3) which constitutes the largest part of the lipid classes in plants [58], [59]. While the specific mechanism of controlling drought-resistance by unsaturation level is still unknown, it is estimated that desaturation attenuates regidification of membranes leading to reduced damage by osmotic stress [60]. The inhibition of beta oxidation could also be interpreted as an effort to decrease the amount of superoxide anion (O_2^-), a toxic by-product reactive oxygen molecule. One of the scavengers of reactive oxygen species, superoxide dismutase (SOD) catalyzes the dismutation of two molecules of superoxide into oxygen and hydrogen [61]. Due to the defensive role of SOD against oxygen toxicity, SOD overexpressed transgenic rice showed enhanced drought or salt tolerance [62], [63]. Likewise, most of the SOD genes is activated in AP2 samples after drought stress (Table

Table 2.2: TRAP pathway analysis results on rice dataset. P_{NDE} and P_{PERT} are the P-value from ORA analysis and SPIA analysis, respectively. The two P-values are combined into P_G by Equation 2.6. $P_{NDE-FDR}$ and P_{G-FDR} are adjusted P-values using false discovery rate. Status of the pathway is denoted as activated or inhibited, which is a relative status in AP2 samples compared to Nip samples. Three pathways were found to be activated in both one time point and the time series analysis result and additionally two pathways were exclusively found only in the time series analysis result, demonstrating the effectiveness of the time series analysis methods.

Time	Pathway Name	P_{NDE}	$P_{NDE-FDR}$	P_{PERT}	P_G	P_{G-FDR}	Status
0 hour	Diterpenoid biosynthesis	0.000	0.010				
1 hour	Plant-pathogen interaction	0.000	0.000	0.200	0.000	**0.001**	Act.
	Plant hormone signal transduction	0.000	0.007	0.093	0.000	0.008	Act.
	Stilbenoid, diarylheptanoid and gingerol biosynthesis	0.000	0.019				
	alpha-Linolenic acid metabolism	0.001	0.033				
6 hour	Diterpenoid biosynthesis	0.000	0.025				
time-series	**Plant hormone signal transduction**	0.000	0.000	0.220	0.000	0.000	Act.
	Diterpenoid biosynthesis	0.000	0.001				
	Biosynthesis of unsaturated fatty acids	0.001	0.022				
	alpha-Linolenic acid metabolism	0.001	0.022				
	Isoquinoline alkaloid biosynthesis	0.002	0.041				

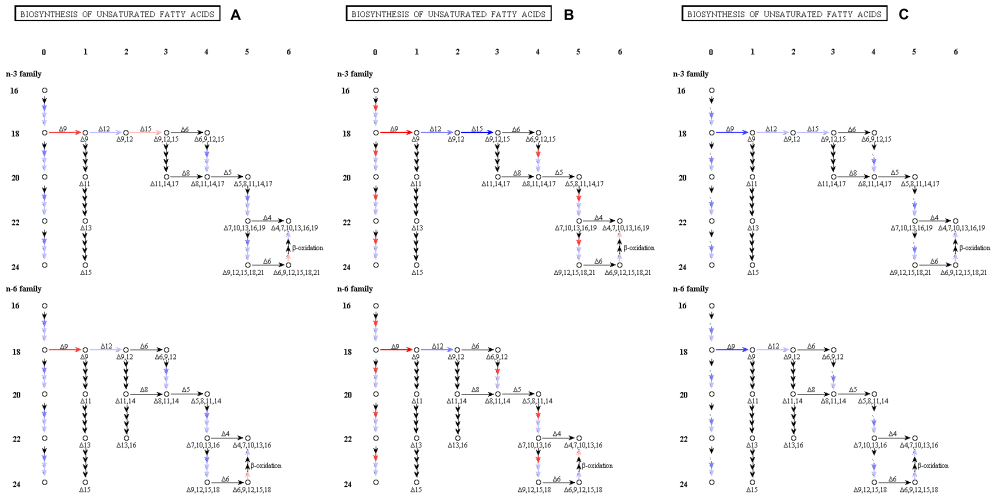


Figure 2.6: Alpha-linolenic acid pathway colored from rice data set. Path related to beta-oxidation is cropped from the original pathway. A, B and C indicates 0, 1, 6 hour after drought stress. Beta-oxidation path is repressed (colored in blue) in 6 hour.

2.3).

The P_{G-FDR} of *Plant-pathogen Interaction* pathway is below 0.05 at 1 hour but not in the time series analysis result due to its increased P_{PERT}, which indicates its signaling impact is much weaker considering the whole time points. Removal of this pathway is reasonable because drought stress is an abiotic stress and its high P_{NDE} might be from the high ratio of common DEGs with Plant hormone signal transduction pathway.

Figure 2.8 is the graphical representation of the pathway analysis result. Only pathways with adjusted P-value above 0.5 are included due to the size

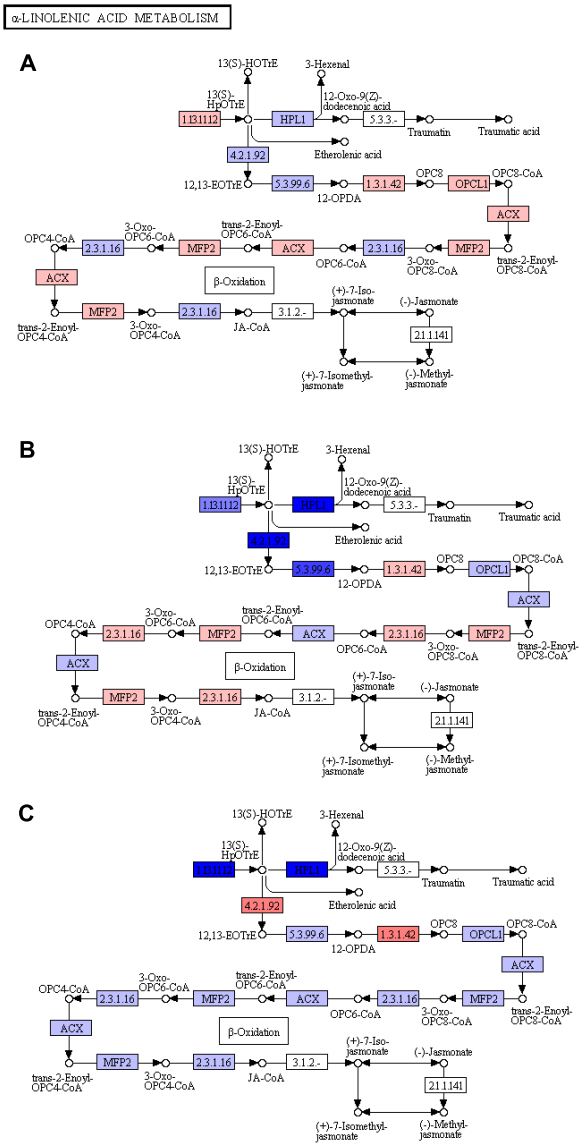


Figure 2.7: Alpha-linolenic acid pathway colored from rice data set. Path related to beta-oxidation is cropped from the original pathway. A, B and C indicates 0, 1, 6 hour after drought stress. Beta-oxidation path is repressed (colored in blue) in 6 hour.

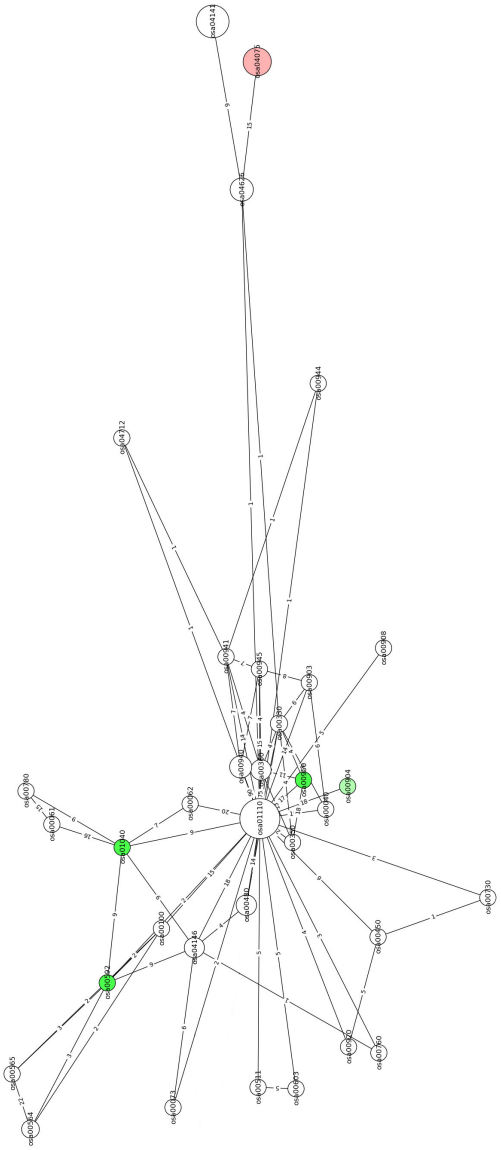


Figure 2.8: The graphical representation of the pathway analysis result including pathways with adjusted P-value above 0.5. Four pathways in green are pathways with unknown status and one pathway in red is *Plant hormone signal transduction* pathway which is found to be activated by SPIA. No genes are found to be common between the five pathways.

of the network. Four pathways in green are pathways with unknown status and one pathway in red is *Plant hormone signal transduction* pathway which is found to be activated by SPIA. Although there is no common gene between five significant pathways, the graph shows the location of significant pathways and clusters in the pathway network.

2.3.2 Clustering results

4 out of 27 possible clusters had significant pathways in the result of clustering analysis with the threshold of 1.0 (Table 2.4). There are two pathways previously found in the pathway analysis using ORA, *Isoquinoline alkaloid biosynthesis* and *Diterpenoid biosynthesis*. However, the status of these pathways (i.e., direction of gene expression change) was unknown since ORA method which estimates the significance of pathways only from the number of DEGs. The estimation of pathway status is possible from the clustering result from the labels of the clusters. *Isoquinoline alkaloid biosynthesis* pathway belongs to cluster DDD which means the pathway is inhibited for the whole time. *Diterpenoid biosynthesis* pathway in cluster DCU seems to be inhibited at 0 hour and activated at 6 hour which is consistent with the pathway analysis result of 0 and 6 hour.

2.3.3 Results from other tools

TRAP result was compared with Gene Set Enrichment Analysis (GSEA) and Significance Analysis of Microarrays (SAM). GSEA is a well-known gene set analysis tool that gives a nominal P-value and a FDR Q-value for each pathway. I used *Log2_ratio_of_classes* as metric for ranking genes and the analysis is done for each time point because GSEA is not originally developed for time series datasets. In the GSEA result (Table 2.5) at 5% cutoff of the FDR Q-value, all of

the significant pathways enriched in Nip samples. In other words, GSEA could not find any pathway activated in drought-resistant rice (Table 2.5) while TRAP found several activated pathways such as *Plant-pathogen interaction* and *Plant hormone signal transduction* pathway. This is because GSEA does not consider the interaction between genes. If two phenotypes A and B are given, GSEA would consider a pathway as enriched in A if most of the genes in the pathway have positive log fold change (when the fold change is calculated as log(A/B)). However, TRAP focuses on the signaling effect of the interaction. Even though there are fewer genes with positive log fold change, if the chain reaction (i.e., gene X activates Y and Y activates Z) exists among them, perturbation factor of each gene increases and the p-value of the pathway becomes smaller. Therefore, TRAP can find more activated pathways in terms of signaling effect. Among pathways that do not have interaction between genes, *Photosynthesis* and *alpha-Linolenic acid metabolism* pathways were detected by both TRAP and GSEA. However, many pathways were only detected by either TRAP or GSEA but not both. This is because TRAP chooses significant pathway by the ratio of DEGs regardless of its expression direction while GSEA evaluates pathways mainly by the ratio of genes expressed in the same direction. For example, *Diterpenoid biosynthesis* pathway found by TRAP includes genes related to the biosynthesis of various types of diterpenoid acids. In *Diterpenoid biosynthesis* pathway, genes generating Gibberellin is down-regulated but genes making other acids are up-regulated. When DEG expression directions are mixed, up and down, GSEA failed to find the pathway as significant.

SAM provides P-values and FDR values of differentially expressed genes from time series array data. I used gene expression values from the rice dataset for the SAM package because the new version of SAM for sequencing data, SAM-seq, does not have a time series analysis option. The output of SAM is a

list of DEGs but does not generate P-value for each pathway, thus it is hard to be compared with the TRAP result. Therefore I used DAVID that calculates P-values for each KEGG pathways from the gene list. SAM found 1586 DEGs and the DEG list was put into DAVID [32] to find significant KEGG pathways. As a result, *Phosphatidylinositol signaling system* pathway had P-value below 0.05 but after FDR correction there was no significant pathway found.

2.4 Conclusion

This study introduced TRAP, a comprehensive package for analysis time series RNA-seq data that performs pathway and clustering analysis. TRAP implements ORA and SPIA methods for analyzing pathways with and without interactions between genes. Besides providing results for each time point, TRAP considers all time points, i.e., time series, and estimates a single pathway-level statistic value using modified versions of ORA and SPIA. Gene expression change labeling used for clustering enables users to estimate the status of the significant pathways found in ORA results. The application of TRAP to the rice dataset showed the effectiveness of the TRAP time series analysis methods to extract biological meanings from time series data when compared to existing tools such as GSEA and SAM. Moreover, TRAP allows various thresholds and user defined pathways composed of genes of interest for analyzing data. This customization is expected to bring aid to the future researches in biology.

Table 2.3: Log fold change level of SOD genes. Positive log fold change indicates genes are up-regulated in AP2 (transgenic drought-resistant rice) samples. Most of the SOD genes are up-regulated in AP2 after drought stress (Bold-faced).

Gene locus ID	0 hour	1 hour	6 hour
Os03g0219200	-0.785	**0.412**	**0.149**
Os03g0351500	0.282	**0.438**	**0.333**
Os05g0323900	-0.245	**0.802**	-0.189
Os06g0115400	0.080	-0.625	-0.848
Os07g0665200	0.279	**0.829**	**0.741**
Os08g0561700	-0.065	0.009	-0.368

Table 2.4: Clustering results on rice dataset.

Clusters	Pathway Name	Pathway Analysis	P_{NDE}	$P_{NDE\text{-}FDR}$	Status
CCD	Photosynthesis	Not Found	0.000	0.000	Inh. in 6 hr
	Photosynthesis - antenna proteins	Not Found	0.000	0.000	Inh. in 6 hr
DDD	Isoquinoline alkaloid biosynthesis	**Found**	0.000	0.006	Inh.
	Tyrosine metabolism	Not Found	0.000	0.017	Inh.
	Pentose and glucuronate interconversions	Not Found	0.000	0.017	Inh.
	Other glycan degradation	Not Found	0.001	0.034	Inh.
DCU	Diterpenoid biosynthesis	**Found**	0.000	0.037	Inh. in 0 hr Act. in 6 hr
UUD	Zeatin biosynthesis	Not Found	0.000	0.048	Act. in 0,1 hr Inh. in 6 hr

Table 2.5: GSEA results on rice dataset. All of the significant pathways were enriched in Nip samples, and no pathway found to be activated in drought-resistant rice. *Photosynthesis* and *alpha-Linolenic acid metabolism* pathways are in common with the TRAP result and two pathways are newly found in GSEA. *Plant-pathogen interaction* pathway which TRAP discarded in the time series result is in the list of significant pathways at 1 hour. *Plant hormone signal transduction* and *Diterpenoid biosynthesis* pathways which passed the ORA and SPIA test in TRAP were not detected by GSEA.

Time	Pathway Name	P-value	FDR Q-value	FWER P-value	Enriched in
1 hour	Plant-pathogen interaction	0.000	0.000	0.000	Nip1
	Photosynthesis - antenna proteins	0.000	0.001	0.002	Nip1
	alpha-Linolenic acid metabolism	0.000	0.012	0.046	Nip1
	Carotenoid biosynthesis	0.002	0.020	0.101	Nip1
	Porphyrin and chlorophyll metabolism	0.000	0.016	0.105	Nip1
	Photosynthesis	0.008	0.017	0.128	Nip1
6 hour	Photosynthesis	0.000	0.000	0.000	Nip6
	Photosynthesis - antenna proteins	0.002	0.033	0.060	Nip6

Chapter 3

Detecting regulators in a network labeled with time series by cross-correlation and influence maximization technique

To understand the dynamic nature of the biological process, it is crucial to identify perturbed pathways in an altered environment and also to infer regulators that trigger the response. Current time series analysis methods, however, are not powerful enough to identify perturbed pathways and regulators simultaneously. Widely used methods include methods to determine gene sets such as differentially expressed genes or gene clusters and these genes sets need to be further interpreted in terms of biological pathways using other tools. Most pathway analysis methods are not designed for time series data and they do not consider gene-gene influence on the time dimension.

In this paper, I propose a novel time series analysis method TimeTP for determining transcription factors regulating pathway perturbation, which narrows

the focus to perturbed sub-pathways and utilizes the gene regulatory network and protein-protein interaction network to locate transcription factors triggering the perturbation. TimeTP first identifies perturbed sub-pathways that propagate the expression changes along the time. Starting points of the perturbed sub-pathways are mapped into the network and the most influential transcription factors are determined by influence maximization technique. The analysis result is visually summarized in TF-Pathway map in time clock. TimeTP was applied to PIK3CA knock-in dataset and found significant sub-pathways and their regulators relevant to the PIP3 signaling pathway.

3.1 Motivation

My goal in this paper is to develop a computational method to perform analysis of time series transcriptome data in terms of biological pathways and also to determine regulators for differentially expressed gene sets or perturbed pathways. Analysis of time series omics data is very difficult and there are only a few tools available [64]. In addition, it is desirable to identify regulators such as TF that are likely to induce changes in transcriptome over time. However, on top of the complexity of analyzing time series data, considering regulators such as TF makes the complexity of the time series data analysis task dramatically high. In this study, I propose a novel bioinformatics method for analyzing time series omics data to identify both perturbed pathways and regulating TFs. Two main ideas are:

1. I start the analysis by identifying perturbed pathways in comparison of control vs. treatment group and then focusing on TFs that are relevant to the perturbed pathways. In this way, much smaller number of TFs and pathways are considered, thus the complexity of the analysis task is

significantly reduced.

2. To systematically analyze the effect of TFs over time, I adopt and further develop the influence maximization technique in the bounded time.

With these two main ideas, I designed and implemented a time series analysis method of finding transcription factors regulating perturbed sub-pathways (TimeTP). The key properties of TimeTP is as follows. (i) TimeTP identifies perturbed sub-pathways that propagate their expression levels along time and also identifies TFs triggering that pathway perturbation by my four-step approach. (ii) TimeTP adopts two well established computational methods, cross-correlation [65] and influence maximization [66], from the fields of signal processing and social network. (iii) The novel framework of TimeTP produces the **TF-Pathway map in time clock** to trace the pathway perturbation triggered from TF to pathway. As well as the effective visualization of TF-Pathway map, TimeTP provides user-friendly interface by handling a diverse range of input data in terms of type of dataset (RNA-seq or microarray) and type of condition (single time series or control-treatment).

The rest of the paper is organized as follows. The process of perturbed sub-pathway mining in TimeTP will be described in Section 3.2.1 to Section 3.2.3. Section 3.2.4 and Section 3.2.5 explains the time bounded network construction and the influence maximization algorithm for finding TFs. TimeTP is tested by using the biological dataset and the result is compared with other pathway/sub-pathway mining tools and regulator analysis algorithms in Section 3.3.

3.2 Methods

The overview of the proposed method is depicted in Figure 3.1. To model pathways over time, I created an augmented pathway graph where a node is a gene

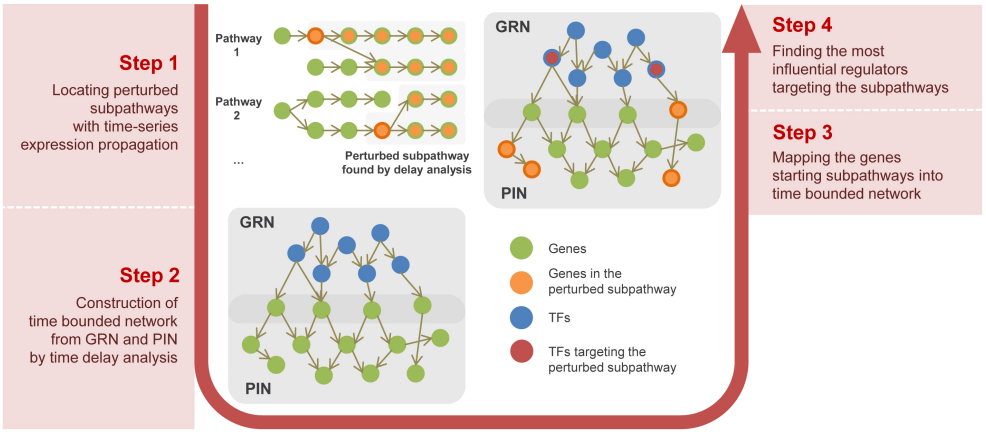


Figure 3.1: Overview of TimeTP analysis workflow. TimeTP uses an integrated network of gene regulatory network(GRN) and protein-protein interaction network(PIN). Each biological pathway is analyzed by TimeTP and the perturbed sub-pathways are identified with the time-delay-bounded propagation of gene expression. The starting point or a gene of each perturbed sub-pathway is mapped to the integrated network and then regulators of the perturbed sub-pathways are identified by the labeled influence maximization algorithm.

augmented with a differential expression vector as an attribute of a node. By measuring cross-correlation, I determine a set of perturbed sub-pathways containing only genes that are connected to propagate expression changes over time. The next major task is to determine TFs regulating the perturbed pathways. In general, TFs are not included in pathways, thus I used gene regulatory network (GRN) to establish orthogonal relationship between regulators and pathways. Identification of regulating TFs requires to estimate the system-wide effect of a TF. To estimate the system-wide influence of a TF, I first augment GRN with protein-protein interaction network(PIN). As a result, I have a network of GRN and PIN combined and the network is big enough to have con-

nections from TF to genes in the pathways. To evaluate of the influence of TF on the perturbed pathways, I used a labeled influence maximization algorithm.

3.2.1 Differential expression vector

Each pathway in the curated pathway database such as KEGG can be represented as a directed graph $G = (N, E)$. Genes and their interactions correspond to nodes and edges in the pathway graph, respectively. TimeTP assigns a time vector $\vec{v}$ for each node, representing the differentially expressed time points as 1(overexpressed) or -1(underexpressed) and otherwise as 0. For example, if data has T number of measured time points and has control and treatment conditions to compare, either -1, 1, or 0, will be assigned for each time point in a differential expression vector $\vec{v}$ of length T. If the data is generated in a single condition, the differential expression can be tested between two time points (e.g. relative to the first time point or adjacent) resulting in a vector of length $T-1$. Whether two groups of samples are differentially expressed is determined by Limma [36] for microarray or by DESeq2 [67] for RNA-seq data.

3.2.2 Perturbed sub-pathway with delay-bounded expression propagation

For each pathway, TimeTP searches for the perturbed sub-pathway by choosing valid edges from the edges in the original pathway information. The validity of edges is determined by looking at the relationship between differential expression vectors of two nodes. I propose two criteria for edges in the perturbed sub-pathway. First, every edge of the perturbed sub-pathway is required to propagate the differential expression pattern along the given direction. Assume that an edge $N_1 \rightarrow N_2$ from a node N_1 to a node N_2 has differential expression vectors $\vec{v_1}$ and $\vec{v_2}$, respectively. The direction of propagation and the number

of delayed time points for a pair of expression vectors can be approximated by cross-correlation, which is a measure of similarity of two time series in signal processing. Cross-correlation of two vectors $\vec{v_1}$ and $\vec{v_2}$ is defined as

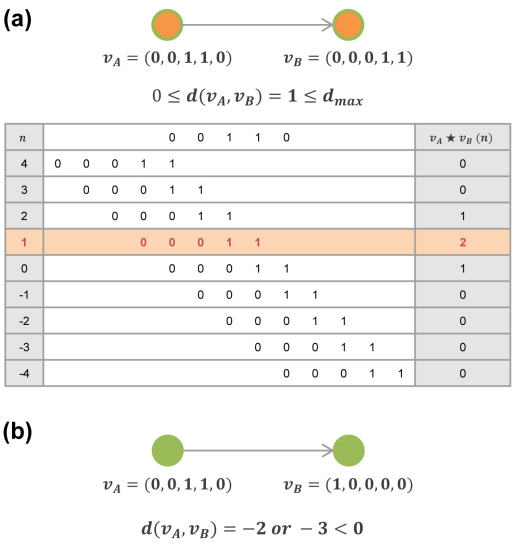
$$(\vec{v_1} \star \vec{v_2})(n) = \sum_{t=-\infty}^{\infty} \vec{v_1}(t)\vec{v_2}(t+n) \tag{3.1}$$

where $\vec{v}(t) = 0$ for $t \leq 0$ or $t > T$ (This happens at the preceding or trailing entries of two vectors). When the two vectors overlap most with n delay, cross-correlation is maximized with a parameter n. Therefore, TimeTP finds the shortest possible delay between two differential expression vectors $d(\vec{v_1}, \vec{v_2})$ where cross-correlation between two vectors is maximized.

$$d(\vec{v_1}, \vec{v_2}) = \underset{n}{\operatorname{argmax}}(\vec{v_1} \star \vec{v_2})(n) \tag{3.2}$$

When $d(\vec{v_1}, \vec{v_2})$ of a directed edge (N_1, N_2) is negative, it implies that the direction of the expression propagation is opposite to the given direction. In this case, the edge is considered as invalid and excluded from the perturbed sub-pathway. Next, a threshold for delay is used to filter out edges with a long positive delay, i.e., bigger than a user defined threshold value, so that the expression propagation in the sub-pathway is bounded within a time period that the user allows. Figure 3.2 shows the examples of delay analysis, where the edge in Figure 3.2(a) has a one time point of delay with maximum cross-correlation 2. Figure 3.2(b) is an example of an invalid edge due to the negative delay. Perturbed sub-pathway with one edge is disregarded. Since TimeTP determines the best delay between two genes, different delays can be assigned to different gene pairs, which can reflect the different speed of signaling steps in the biological pathways.

Once perturbed sub-pathways with bounded propagation is determined from each pathway, source nodes with no incoming edge in the sub-pathways



n												$v_A \star v_B\,(n)$
4	0	0	0	1	1							0
3		0	0	0	1	1						0
2			0	0	0	1	1					1
1				0	0	0	1	1				2
0					0	0	0	1	1			1
-1						0	0	0	1	1		0
-2							0	0	0	1	1	0
-3								0	0	0	1	1	0
-4									0	0	0	1	1	0

Figure 3.2: Cross-correlation examples. (a) Cross-correlation of two vectors $\vec{v_1}$ and $\vec{v_2}$ is maximized with the delay 1. The directed edge is valid and remains in the graph, because the estimated delay is non-negative. (b) Cross-correlation of two vectors $\vec{v_1}$ and $\vec{v_2}$ is maximized with the delay -2 or -3, which indicates the optimal direction of the edge is opposite. The edge is invalid and removed from the graph.

are labeled as targets in the time bounded network. Node weights of labeled source nodes are set as the number of nodes in the sub-pathway and for the other nodes not labeled, zero or negative numbers are assigned so that no profit can be earned from non-labeled nodes. This profit assignment scheme is used to define and rank regulators.

3.2.3 P-value for perturbed sub-pathway

P-value of the perturbed sub-pathway is estimated by the permutation test. The null hypothesis is that perturbed sub-pathways determined by TimeTP are

Algorithm 1 Greedy Labeled IM (G, k)

Initialize $S = \phi$, $SeedInfSet = \phi$, $T = t | t \in TFSet$ and $Round = 1000$

for $n \leftarrow 1, k$ **do**

    Set $Inf[t] = 0$, for all $t \in T$.

    for $i \leftarrow 1, Round$ **do**

        Derive G' by removing each edge from G according to the edge probability $1 - p$.

        for each node $t \in T \setminus SeedInfSet$ **do**

            $$Inf[t] \leftarrow Inf[t] + \sum_{v \in InfSet(t,G')} Profit(v)/len(InfSet(t, G'))$$

        end for

    end for

    $newSeed \leftarrow t \in T$ with maximum $Inf[t]/Round$

    $S \leftarrow S \cup \{newSeed\}$

    $SeedInfSet \leftarrow SeedInfSet \cup InfSet(newSeed)$

    $T \leftarrow T \setminus \{newSeed\}$

end for

randomly generated, without considering the order of genes and their expression patterns in the pathway. To test the hypothesis, differential expression vector for each gene is randomly re-assigned from the vector set of the whole genes and sub-pathways are sampled according to the same procedure described in section 3.2.2. Given that the ratio of DEGs is not aberrantly high, sampled sub-pathways determined from the randomly assigned expression vectors are most likely to have a short path length. Cross-correlation of each edge is likely to be small as well due to the short overlap of two expression vectors. Therefore, a sum of the cross-correlation of every node pair in the sampled sub-pathway is chosen to be a pathway-level statistic and the p-value for a perturbed sub-pathway is derived as the probability of having higher statistics in the sample distribution.

3.2.4 Time bounded network construction

To search for upstream regulators of perturbed sub-pathways, an integrated network of GRN and PIN is constructed. Interaction information between transcription factors (TFs) and target genes (TGs) in GRN is derived from HTRIdb [68] that provides experimentally verified or computationally predicted TF-DNA binding sites from six public databases and literature [69]. Protein-protein interaction for PIN is from STRING [70] database. Integrated network of GRN and PIN is used to determine TFs that have the most overall effect on perturbed sub-pathways and to connect the TFs and perturbed sub-pathways. Two expression vectors that do not preserve the time order are filtered out so that expression propagation along the connecting path is always valid in terms of time clock, as described in Section 3.2.2. This process produces a time bounded network. As for undirected edges of PIN, delay of Equation 3.2 is calculated for both directions and directed edge with nonnegative delay remains.

3.2.5 Labeled influence maximization for transcription factor detection

The main goal of influence maximization is to locate a set of seed nodes in the network that could maximize the spread of influence [66] and the technique has been successfully used to select marketing targets in the social network. A modified version called the labeled influence maximization developed by Li et al. [71] exploits profit values of nodes to prefer seed nodes that have an influence on a specific node set. TimeTP utilizes a greedy version of the labeled influence maximization algorithm to the time bounded network with a few modifications (see below) so that influence of a gene on the perturbed sub-pathways are properly modeled.

Labeled influence maximization algorithm (Algorithm 1) used in TimeTP is intended to determine the most influential k regulators in the time bounded network G, especially TFs targeting the starting nodes of the perturbed sub-pathways. It first initializes a set of seed nodes S and a set of nodes that can be influenced by seed nodes $SeedInfSet$ as an empty set. For every TF t not selected as a seed node and not influenced by the current seed nodes, its influence $Inf[t]$ is quantified by the average profit values of nodes that the TF can reach. The same procedure is repeated for *Round* times creating a subgraph G' from G according to the edge weight between 0 and 1 regarded as a probability of an edge (line 4-9). Probability of edges are derived from the confidence score of STRING and 1 for GRN edges. After the iteration, a TF with the maximum influence is included in the seed set and $SeedInfSet$ is updated as well.

Table 3.1: Significantly perturbed pathways in PIK3CA H1047R samples found by TimeTP and comparison with other representative pathway tools (+: found, -: not found). Pathways with p-value below 0.05 are shown. (TS:time series)

Pathway	Pathway name	DEG p-value	Path	Path p-value	Combined p-value	Sub-pathway		Pathway		Ref.
						non-TS	TS	non-TS	TS	
						DEAP	TimeClip	SPIA	TRAP	
hsa04012	ErbB signaling pathway	0.000	Path1	0.000	0.000	-	-	-	-	[72]
hsa04810	Regulation of actin cytoskelet...	0.000	Path1	0.001	0.000	-	-	-	-	[73]
			Path2	0.005	0.000					
hsa04520	Adherens junction	0.000	Path1	0.020	0.000	-	-	-	-	[74]
										[75]
hsa04310	Wnt signaling pathway	0.001	Path1	0.021	0.000	-	+	-	-	[76]
										[77]
hsa04510	Focal adhesion	0.000	Path1	0.024	0.000	-	+	+	+	
hsa04068	FoxO signaling pathway	0.004	Path1	0.027	0.000	-	-	-	-	[78]
hsa04666	Fc gamma R-mediated phagocytos...	0.003	Path1	0.019	0.001	-	-	-	-	[79]
hsa04151	PI3K-Akt signaling pathway	0.032	Path1	0.005	0.001	-	+	-	+	[80]
hsa04114	Oocyte meiosis	0.032	Path1	0.020	0.005	-	-	-	-	[81]
hsa04921	Oxytocin signaling pathway	0.032	Path1	0.023	0.006	-	-	-	+	[81]

3.3 Result

TimeTP is tested with a genome-wide RNA-seq dataset of non-transformed human breast epithelial cells MCF10a starved overnight and stimulated with 10 ng/ml EGF for 15, 40, 90, 180 and 300min [80], in WT and PIK3CA knock-in samples. To test the power of influence maximization, I need to choose datasets with many time points and also with sequencing data to accurately model influence of TFs. Note that many datasets with only 2 time points are not meaningful for this analysis. In addition, to test the performance of the proposed approach, data should have replicates to determine differential expression accurately and the interval between time points should be short to model signaling effects. The MCF10a data was the only one to meet the criteria.

PIK3CA knock-in samples (referred to as 'PIK3CA H1047R') contain a mutated gene that encodes the p110α catalytic subunit (PIK3CA). PIK3CA is a component gene of Class IA phosphoinositide-3-kinases (PI3Ks) and the

mutated form of PIK3CA is expected to exhibit chronic activation of phosphatidylinositol (3,4,5)-trisphosphate (PIP3) signalling. PI3K/PIP3 signaling pathway plays a key role in cell growth and migration. In addition, several driver mutations in PI3K/PIP3 pathway have been found in multiple types of cancer. Especially, oncogenic mutations of PIK3CA gene are discovered in up to 45% of human breast cancer samples [82]. Thus, this experiment is designed to trigger long term activation of PIP3 signaling by the modification of PIK3CA and track its downstream effect. Analysis result of TimeTP is composed of the TF-Pathway map in time clock and the whole list of perturbed pathways as shown in Figure 3.3 and Table 3.1. Javascript library of Cytoscape is used for TF-Pathway map visualization [83].

3.3.1 TF-Pathway map in time clock

Figure 3.3 is the map of influence path from the transcription factors selected by the influence maximization algorithm to perturbed sub-pathways. Pathways perturbed but not affected by transcription factors are excluded in the TF-Pathway map. For example, TimeTP detected perturbation of the PI3K-Akt signaling pathway (Table 3.1) but it was not included the TF-Pathway map in time clock (Figure 3.3) because PI3K-Akt signaling pathway is directly activated by the modification of PIK3CA in the experiment.

As in Figure 3.3, **FOXO4** is on the top of the TF-Pathway map and propagates its effect to all of the downstream pathways and FoxO signaling pathway itself. The forkhead box O (FoxO) transcription factors are known as targets of the serine/threonine protein kinases (PKB)/Akt [78] that is directly affected by PIP3 generation [80]. Specifically, Akt inhibits FoxOs and causes consequent inactivation of **FoxO signaling pathway**, which can be clearly shown in the TF-Pathway map of PIK3CA H1047R samples. **Wnt signaling path-**

way is one of the activated pathways in PIK3CA H1047R samples. TimeTP estimated that differential expression of the transcription factor **FOXO4** and **SREBF1** in the early time points (1 to 3) is propagated through the path and activated Wnt signaling. Although the first gene GSK3B of the perturbed sub-pathway is down-regulated, consequently it made CTNNB1 that encodes β-catenin activated to further transduce the signal to other cytoplasmic regions or into the nucleus. Interaction of FoxOs with β-catenin has an inhibitory effect on β-catenin activity [75], while TimeTP inferred a devious route that has the same consequence. As for the cooperation between PI3K-Akt signaling and Wnt signaling, several studies provide the logical underpinnings [76, 77].

The activation process of **ErbB signaling pathway** and **Regulation of actin cytoskeleton** is more complicated. Albeit both perturbed sub-pathways themselves are seemingly down-regulated first, the path from TF to the first genes of the sub-pathway (ACTB, ACTG1, HBEGF) is activated and finally three genes are activated in the last time point, forecasting the late activation of two pathways beyond the observed time points. As in the previous studies [72], ErbB signaling pathway encompasses the PI3K-Akt signaling pathway. The perturbed sub-pathway that TimeTP detected in the ErbB signaling pathway includes the cell surface receptor EGFR, which is the upstream part of PI3K-Akt pathway. Taken together, the effect of Akt signaling activation attributes the delayed activation of the ErbB signaling, which can be the positive feedback loop of the Akt signaling pathway. Detection of a transcriptional feedback loop of PIP3 signaling is the major contribution of the original paper of the dataset [80] and the analysis result of TimeTP can be a parallel contribution of the study. Moreover, TF-Pathway path to the Regulation of actin cytoskeleton pathway found in TimeTP result suggests for further research in addition to the previously suggested path [73].

Fc gamma-R mediated phagocytosis is one of the activated pathways in PIK3CA knock-in samples. Mammary epithelial cells can act as phagocytes [84]. During phagocytosis, ligated Fc gamma-R on plasma membranes induces recruitment of PI3K and increased synthesis of PIP3 [79]. TimeTP found the perturbation of phagocytosis pathway starting from PI3K receptor and its downstream genes. One of the TFs expected to trigger the perturbation is **ATF3** (Figure 3.3) down-regulated in the early time points, which is a key regulator that inhibits the immune response of macrophage [85]. My analysis correctly suggested that ATF3 would function similarly in MCF10a cells. TimeTP detected the same sub-pathway in **Oocyte meiosis** and **Oxytoxin signaling**. In both pathways, Calcium/calmodulin (CALM) signaling pathway is included and its sub-pathway was found as perturbed. Previous studies of mammary carcinoma cells report that calmodulin mediates Akt activity [81, 86], suggesting that the increased PIP3 not only recruited Akt by itself but also induced calmodulin-dependent activation of Akt signaling pathway.

3.3.2 Comparison with existing pathway/regulator analysis tools

Most of the pathway analysis tools assume that the expression value for each gene follows the Gaussian distribution, which is not appropriate next generation sequencing data. Therefore, I selected four representative tools without the Gaussian assumption in each class of pathway analysis tools: DEAP(sub-pathway analysis), timeClip(sub-pathway analysis, time series), SPIA(pathway analysis), TRAP(pathway analysis, time series). Samples of different time points are treated as replicates in DEAP and SPIA that do not perform time series analysis, and WT samples are not used for TimeClip that does not support control vs. treatment group analysis. Table 3.1 shows a list of sub-pathways with significant expression propagation from TimeTP analysis. DEAP and SPIA

Table 3.2: Relevance of pathways detected by TimeTP and other pathway analysis tools to PI3K. Relevance to PI3K is evaluated using a state of the art context-aware literature search tool, BEST [87].

	Avg. Score of PI3K
TimeTP	674.109
DEAP	000.000
TimeClip	453.296
SPIA	267.412
TRAP	271.771

failed to choose most of the pathways including PI3K-Akt signaling pathway that is expected to be activated in PIK3CA H1047R samples, presumably due to the disregard of time factor. timeClip and TRAP selected out more significant (sub-)pathways, yet disregarded FoxO signaling pathway and ErbB signaling pathway presumably pertaining to PIP3 signaling as described in Section 3.3.1.

For the pathways that are not detected by TimeTP, we performed literature based analysis to evaluate how these pathways are relevant to the PI3K knock-in condition. Since PI3K is extensively studied, there are so many studies related to PI3K, so a simple literature search on Pubmed produced a huge number of articles that cannot be compiled. Thus, we used a state of the art context-aware literature based search engine, BEST [87]. The average BEST scores from http://best.korea.ac.kr are summarized in Table 3.2. Thus, we show that pathways detected by TimeTP are not only confirmed by references but also by a contextual literature search tool, BEST –context is relationship from the query, PI3K in this case.

Running times of the methods are compared in Table 3.3. Even though

Table 3.3: Running time of TimeTP and other pathway analysis tools using MCF10a dataset with 6 time points. Different from other tools, TimeTP include an additional process for regulator analysis in addition to pathway analysis. The number in brackets is the running time measured for pathway analysis only. (Intel(R) Xeon(R) CPU E7-4850 2.00GHz, CentOS 5.8. Python 2.7.3. for TimeTP, DEAP and TRAP. R 3.2.2 for timeClip and SPIA.)

Tool	TimeTP	DEAP	timeClip	SPIA	TRAP
Time (s)	906.767 (408.862)	1776.885	94.989	398.474	1125.961

TimeTP performs an additional step, that is, influence maximization, compared to competing pathway analysis tools, overall running time of TimeTP is similar to those of other pathway analysis tools. As for the sub-pathway detection process of TimeTP, running time of TimeTP (408.862s) is smaller than the average running time of other four tools (849.077s) and the overall process including the regulator search by influence maximization takes similar time (906.767s) in average.

Two regulator analysis methods are compared with TimeTP. MRA is a method for selecting and ranking TFs in gene regulatory network and DREM is a tool for time series clustering. WT samples are not used for DREM that does not support control vs. treatment analysis. Table 3.4 is the list of master TFs selected from TimeTP, MRA and DREM. TFs from TimeTP are regulators of the perturbed sub-pathways chosen and ranked by the labeled influence maximization algorithm. Among 16 TFs from the TimeTP result, USF1, TGIF1 and RREB1 are TFs expected to bind to strongly genes regulated in the PI3K signaling-activated samples based on the motif activity analysis in the original

paper of the dataset, corroborating the credibility of TimeTP. MRA performs Fisher's exact test to first confirm the ratio of signature genes among its target genes and ranks TFs that passed the test by the number of signature genes. To apply the same standard with TimeTP, TF-gene interaction information is extracted from the same GRN and 18 genes that starts the perturbation in each sub-pathway are used as signature genes for MRA. However, only one TF, SREBF1, that directly targets Wnt signaling pathway satisfied the criteria of MRA. Distinct from the one-to-one mapping of a MRA, the influence maximization algorithm rescued 15 TFs with indirect influence on targeted pathways in the network structure in addition to SREBF1. Furthermore, TFs that target multiple pathways are prioritized higher than TFs with a single target. LMO2, ATF3, FOXO4 and RFX1 in the Figure 3.3 are such examples. TFs that do not target multiple pathways but are highly ranked have the small number of downstream genes, thus the ratio of genes in the perturbed pathway is relatively high among its downstream genes. DREM performed time series clustering and found three TFs different from TimeTP or MRA, regulating one of the clusters (Figure 3.4). The three TFs target the same cluster with 71 genes, but the cluster is not enriched with any KEGG pathway by Fisher's exact test (Table 3.5).

3.4 Conclusion

I presented TimeTP, a four-step approach to locate perturbed sub-pathways and their regulators from time series transcriptome data. TimeTP has two novel contributions: estimation of delay between two expression vectors that leads to the construction of a time bounded subgraph, and introduction of

Table 3.4: Transcription factors found by TimeTP and other tools. TFs in boldface are the intersection with TFs selected as significant in the original paper.MRA detected only one transcription factor and DREM does not provide ranks of the detected TFs.

TimeTP		MRA		DREM
Rank	TF	Rank	TF	TF
1	NKX3-1	1	SREBF1	FOXF2, NF1, SRF
2	LMO2			
3	ATF3			
4	FOXA1			
5	CEBPA			
6	FOXO4			
7	FOXL1			
8	RFX1			
9	**TGIF1**			
10	SREBF1			
11	FOXO3			
12	USF2			
13	**USF1**			
14	GTF2A1			
15	RORA			
16	**RREB1**			

Table 3.5: Clustering result of DREM. Pathway analysis (Fisher's exact test) is performed by DAVID [32].

Cluster	# of genes	Pathway	P-value	TF
Gray	71	NA	NA	FOXF2, NF1, SRF
Green	132	Focal adhesion	4.30E-04	
		Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5.10E-03	
		Regulation of actin cytoskeleton	1.30E-02	
		Leukocyte transendothelial migration	2.30E-02	
		Pathways in cancer	2.90E-02	
		Small cell lung cancer	4.20E-02	
		ECM-receptor interaction	4.20E-02	
		Hypertrophic cardiomyopathy (HCM)	4.30E-02	
		Hematopoietic cell lineage	4.40E-02	
Brown	197	Phosphatidylinositol signaling system	2.60E-02	
		Taste transduction	7.40E-02	
		Inositol phosphate metabolism	7.90E-02	

the influence maximization technique into the analysis of times series data in search of TFs that are involved in perturbed pathways. TimeTP is the first sub-pathway mining tool for time series data that analyzes and visualizes the explicit expression pattern, providing a holistic picture of the pathway perturbation dynamics. In particular, TF-Pathway map in time clock enables user to navigate the perturbation propagation route along time.

Analysis of the PIK3CA knock-in dataset shows that TimeTP can capture the perturbation in PI3K-Akt signaling, confirming the main objective of the biological experiment and re-producing consequent changes in the downstream pathways. Especially, FOXO4 is expected to be the master regulator of the perturbation of five pathways in TF-Pathway map, which is in an agreement with the fact that FoxO transcription factors are the known targets of Akt. As well as the perturbation in FoxO and Wnt signaling pathway directly affected by FoxOs, TimeTP suggests the late activation of ErbB pathway that highlights the same assumption of previous study, a positive feedback loop of the Akt signaling. In addition, TFs predicted and ranked by TimeTP include three important TFs from the original paper of the dataset while MRA or DREM failed to discover any TF in the list.

TimeTP supports various types of dataset with flexible parameters that can be adjusted for the search of regulators. I believe that TimeTP will be a very valuable tool to identify both perturbed pathways and their regulators, especially in analysis of time series sequencing data.

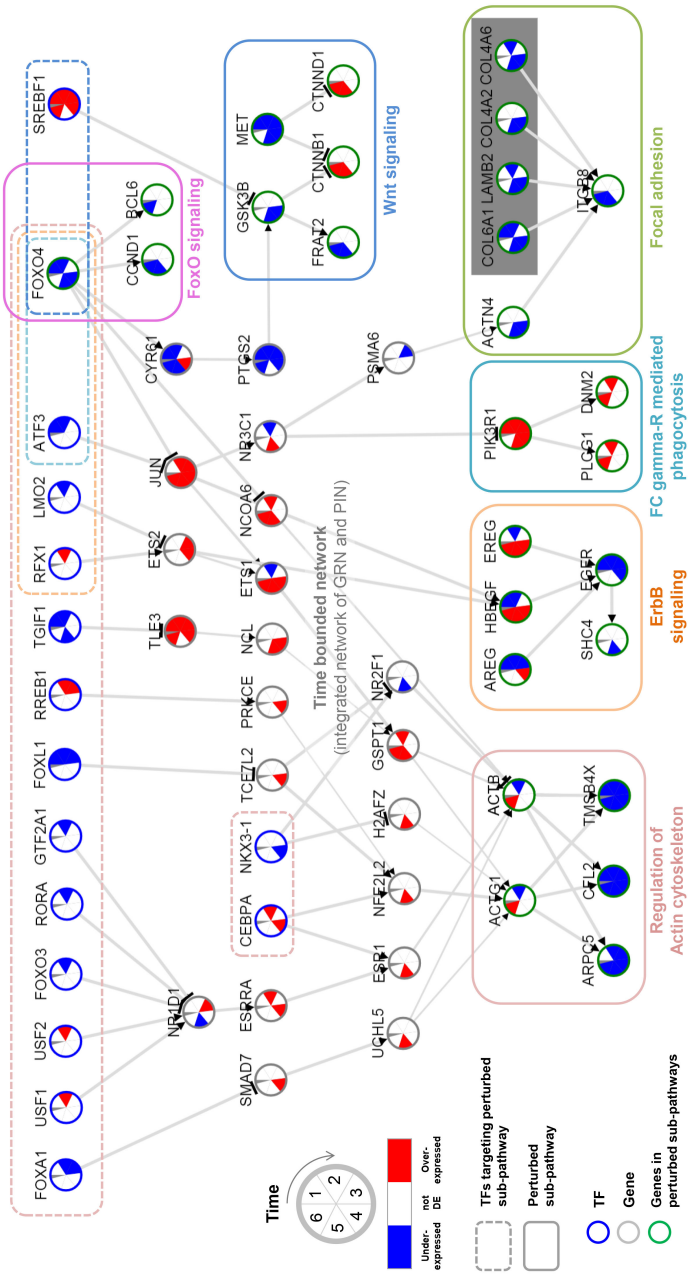


Figure 3.3: TF-Pathway map in time clock. Expression propagation paths from TFs to perturbed sub-pathway genes in the integrated network are specified. For each node, time series differential expressions are colored in a clockwise direction.

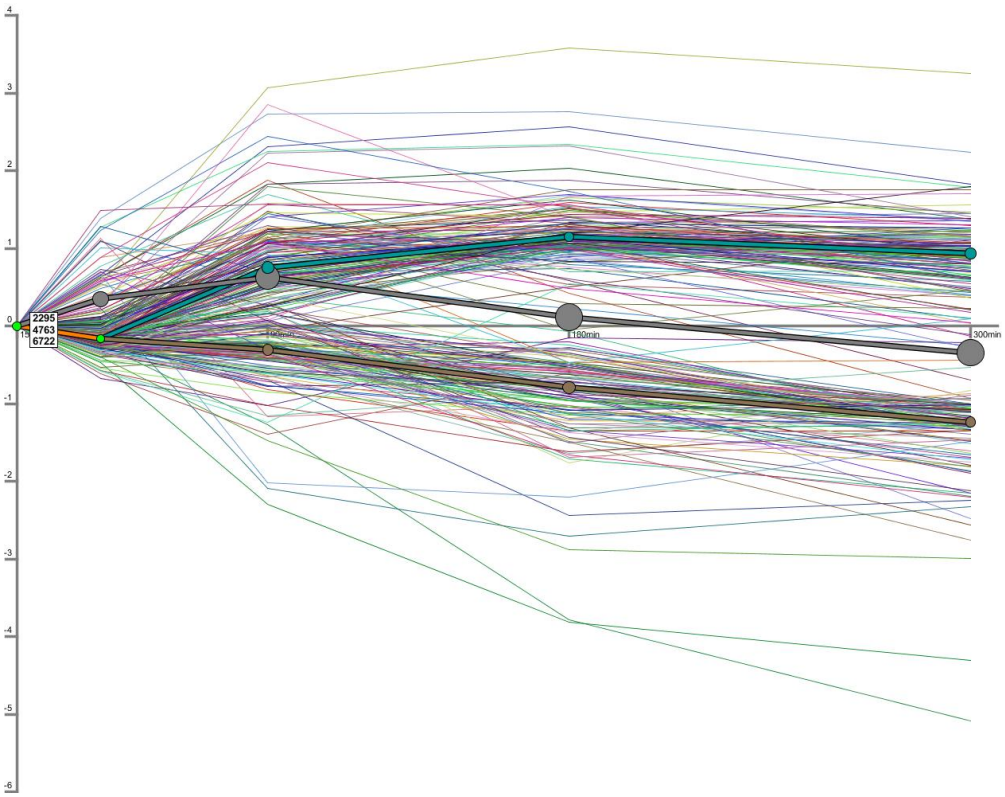


Figure 3.4: DREM result. DREM found three clusters (green, gray and brown lines from the top). Three TFs are predicted to target a cluster colored in gray. Gene symbols (entrez IDs) of the TFs are FOXF2 (2295), NF1 (4763), SRF (6722).

Chapter 4

Inference of cluster network from unevenly spaced time series by Gaussian process and shape-based clustering

Clustering genes with similar expression pattern is one of the most common approaches for analyzing gene expression time series. Most of the algorithms developed so far assume evenly spaced time series and do not take the time interval into account. However, biological experiments are often performed in unevenly spaced time intervals due to the experimental constraints on biological samples. This study aims to incorporate Gaussian process regression into the clustering process to predict unseen values of time series and provide more accurate clustering result. In addition, a network of clusters is generated by measuring distance of expression patterns between clusters. As a distance measure, shape-based distance (SBD) will be applied to capture similarity between time-shifted patterns. Integrating SBD into the construction of cluster network

is expected to infer gene regulatory relationship between clusters considering cascading signaling effect over time.

4.1 Introduction

Selecting the appropriate sampling rate of time points for biological experiment is difficult because the underlying principle of biological process has not been elucidated thoroughly. The most common approach of the biological experiments for developmental and response processes is using the decreasing sampling rate (Bar-Joseph et al., 2012). It is because most of the transcriptional response occurs at the early time points. Figure 11 shows an example of the sampling rates in biological experiment. Samples are captured with decreasing sampling rates at 0, 15, 40, 90, 180 and 300 minutes after the stimulation. Samples in late time points are used to differentiate between transient and sustained response compared with the early samples.

As for clustering gene expression time series, several algorithms have been developed. However, most of the algorithms do not consider the uneven sampling rates explained in Chapter 4.1.1 and uses the time series as it is sampled. Clustering algorithms that are commonly used in the biological research papers such as K-means and hierarchical clustering are also based on the assumption that the time points are even and independent. In addition, following analysis after clustering genes is limited to the functional analysis of each cluster. It can ignore the relationship between genes in different clusters which can share a common function but are divided into different clusters due to the response time delay.

This study aims to solve the problem of clustering time series in biological data described in the previous section. I suggest a framework to generate

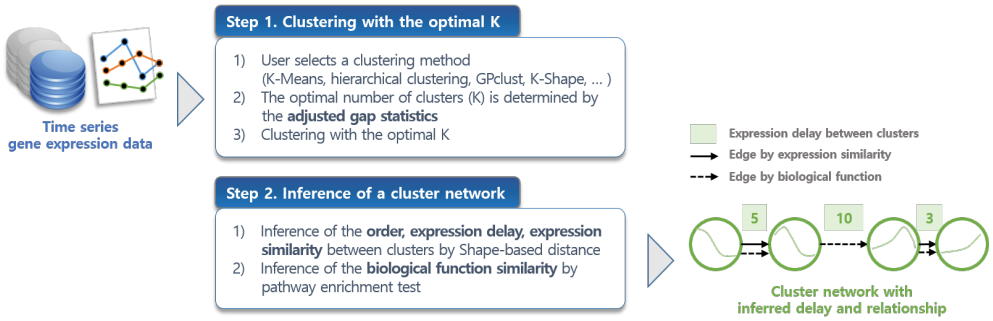


Figure 4.1: Overview of TiClNet

Time series Cluster Network (TiClNet) from gene expression time series. The following sections introduce two steps of TiClNet to improve clustering results by Gaussian process and Shape-Based Clustering:

1. Calculating gap statistics using distance measure for Gaussian process to find the optimal number of clusters

2. Construction of the cluster network using the edge definition by Shape-Based Distance (SBD)

4.2 Methods

The overview of the proposed method is described in Figure 4.1. Users can select an algorithm for clustering including K-means, agglomerative clustering (with different linkage criteria) and K-Shape. The optimal number of cluster (K) is determined by the adjusted gap statistics for time series. After clustering with the optimal K, a network of cluster is generated and visualized to track the dynamic expression patterns of clusters and their relationships.

4.2.1 Gap statistics using distance measure for Gaussian process

Gap statistics is a method for estimating the number of clusters (k) in a set of data [88]. The algorithm calculates gap statistics by comparing the change in within-cluster dispersion with that expected under a reference null distribution for each k in a given range 1, ..., K. The optimal number of clusters is determined by the gap between two within-cluster dispersion from observed and reference dataset. k can be where the gap is maximized or the minimum k where the gap is within the standard deviation of the global maximum gap [89]. As a measure of variance within clusters, the normalized intra-cluster sums of squared Euclidean distance called variance quantity W_k is calculated as the sum of the pairwise distances for all points in cluster C_k, where μ_k is the mean of data points x in the cluster and n_k is the number of data points in the cluster:

$$D_k = \sum_{x^i \in C_k} \sum_{x^j \in C_k} ||x^i - x^j||^2 = 2n_k \sum_{x^i \in C_k} ||x^i - \mu_k||^2 \tag{4.1}$$

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k \tag{4.2}$$

Gap statistics is calculated by comparing $log(W_k)$ with that of a null reference distribution of the data, i.e. a distribution with no obvious clustering with the number of data points n.

$$Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k) \tag{4.3}$$

Assuming each gene expression is generated from a Gaussian process, the distance measure used in calculating variance quality can be changed to consider time dependency. Rather than calculating the Euclidean distance between a data point and cluster mean, I can use the likelihood of a data point from a

Gaussian process of the mean Likelihood of a data point is calculated as the sum of the probability that the observed data point is generated from the Gaussian process of the cluster mean.

$$L_k = \sum_{x^i \in C_k} \sum_{x^j \in C_k} ||x^i - x^j||^2 = 2n_k \sum_{x^i \in C_k} ||x^i - \mu_k||^2 \tag{4.4}$$

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k \tag{4.5}$$

Contrary to the Euclidean distance, the larger likelihood indicates smaller within-cluster variance. Therefore, likelihood of reference dataset is subtracted from that of observed dataset to find the optimal K where the gap is maximized.

$$Gap_n(k) = log(W_k) - E_n^*\{log(W_k)\} \tag{4.6}$$

4.2.2 Edge definition by Shape-Based Distance (SBD)

Shape-Based Distance (SBD) is a distance metric developed for time series considering the shifted pattern of two series. Among all possible shifts $w \in \{1-m, m-1\}$ between two sequences, SBD uses cross-correlation (CC) of the sequences at the optimal shift where the cross-correlation is maximized. In this study, SBD is used to define an edge between two clusters a and b with their expected means μ^a, μ^b:

$$SBD(a,b) = 1 - \max_w \frac{CC_w(\mu^a, \mu^b)}{\sqrt{R_o(\mu^a, \mu^b) R_o(\mu^a, \mu^b)}} \tag{4.7}$$

Calculating SBD between every pair of clusters, I can derive cluster-cluster distance (similarity) and their expression shift. SBD between two clusters is compared with the reference null distribution H_0 of SBDs from random pairs of sequences so that their significance $P(a,b)$ is drawn as the probability of less

values observed in H_0 than $SBD(a, b)$. The optimal shift w inferred from SBD is used to order two clusters. I assume cluster a is followed by cluster b when $w < 0$ and vice versa.

4.3 Results

[90] measured genome-wide mRNA transcript levels during the cell cycle of the budding yeast S. cerevisiae. Cdc28-13 cells were collected at 17 time points taken at 10 min intervals, covering nearly two cell cycles. The authors defined 220 functionally characterized genes for each cell cycle phase (early G1, late G1, S, G2, M phase). Dataset from this study has been used as benchmark for clustering time series as there are few datasets with cluster labels for time series.

4.3.1 The optimal number of clusters from gap statistics

To verify the accuracy of the adjusted gap statistics using Gaussian process likelihood, the optimal number of clusters inferred by TiClNet is compared with those from the original gap statistics using Euclidean distance. Clustering is performed with four different algorithms (Table 4.1) with the given K ranging from 1 to 20 and gap statistics are calculated with Gaussian process likelihood and Euclidean distance for each k. K-means and agglomerative clustering (using ward's and complete linkage) are selected as representatives of the classic clustering algorithms not considering time dependency. K-shape [91] is an algorithm for clustering time series with Shape-Based Distance, even capturing the similarity of shifted sequences. Additionally, Bayesian hierarchical clustering (BHC) [92] is included to the comparison. BHC is based on agglomerative clustering using Gaussian process likelihood for merging criteria. Gap statistics

Table 4.1: Clustering algorithms used to evaluate the accuracy of the adjusted gap statistics.

Algorithm	Abbrebiation	Time dependency	Shift	Automatically detect K	Description	Ref.
K-means	KM	X	X	X		[93]
Agglomerative clustering	AC (w)	X	X	X	Linkage: ward's criteria	[94]
Agglomerative clustering	AC (c)	X	X	X	Linkage: complete linkage	[95]
K-shape	KS	O	O	X		[91]
Bayesian hierarchical clustering	BHC	O	X	O		[92]

Table 4.2: Optimal number of cluster (k) inferred from gap statistics and clustering performance evaluation with inferred k. (F1: F1 score, ARI: Adjusted rand score, Sil.: Silhouette score)

Algorithm	Gap statistics using GP				Gap statistics using ED			
	Optimal k	F1	ARI	Sil.	Optimal K	F1	ARI	Sil.
K-Means	7	0.536	0.410	0.216	18	0.239	0.158	0.100
AC(w)	7	0.559	0.428	0.218	17	0.353	0.260	0.130
AC(c)	12	0.556	0.427	0.181	16	0.399	0.281	0.119
K-Shape	16	0.231	0.134	0.015	19	0.213	0.131	-0.045
	Optimal K	F1	ARI	Sil.				
BHC	40	0.119	0.065	-0.156				

is not calculated for the result of BHC because it automatically detects the optimal number of clusters.

Table 4.2 shows the optimal number of cluster (k) inferred from gap statistics and clustering performance evaluation with inferred k. Adjusted gap statistics using Gaussian process predicts more accurate number of clusters (true k=5) than Euclidean distance for all clustering algorithms tested. Consequently, performance evaluation scores (F1 score, adjusted rand index, silhouette score) are higher with the ks from TiClNet closer to 5. BHC inferred much higher number of clusters than the expected k than those inferred from gap statistics, thereby achieving lower performance.

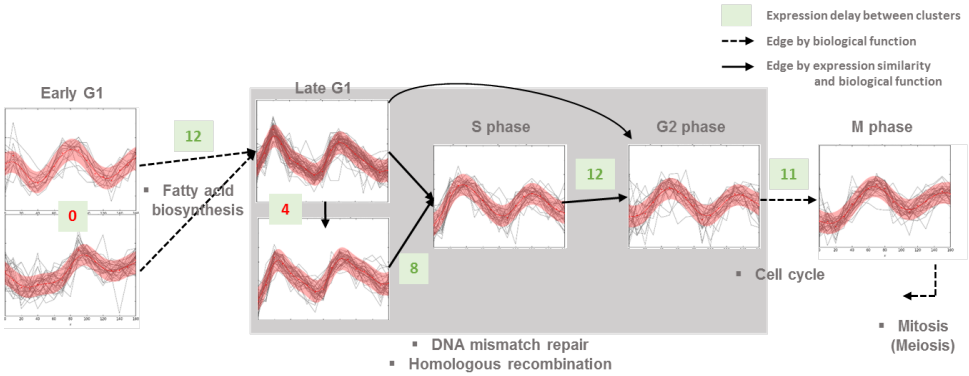


Figure 4.2: Cluster network inferred by TiClNet.

4.3.2 Cluster network of cell cycle dataset

The final result of this study is the cluster network with clusters ordered by their expression profiles and edges derived from the relationship between the clusters. Assuming a user selected K-means algorithm and TiClNet suggested the number of clusters as 7, a network of 7 clusters is generated as Figure 4.2.

Each graph shows an expression pattern of the cluster and expression shifts (delays) between clusters are specified in green boxes. There are two pairs of clusters with the smallest delay (0 and 4, red texts in green boxes). According to the true label of these clusters, one pair of clusters with 0 delay belongs to the early G1 phase and the other pair with 4 delay to the late G1 phase. This shows that even though the number of clusters inferred by TiClNet is larger than 5, TiClNet is able to group similar clusters by the small shift between them. In addition, edges derived from the expression similarity between clusters (solid arrows) detected a group of DNA repair-relevant clusters with late G1, S, G2 phase genes (gray box). Biological functions enriched for 5 edges of these clusters all included DNA mismatch repair or homologous recombination pathway that is essential for the accurate repair of DNA double-strand breaks.

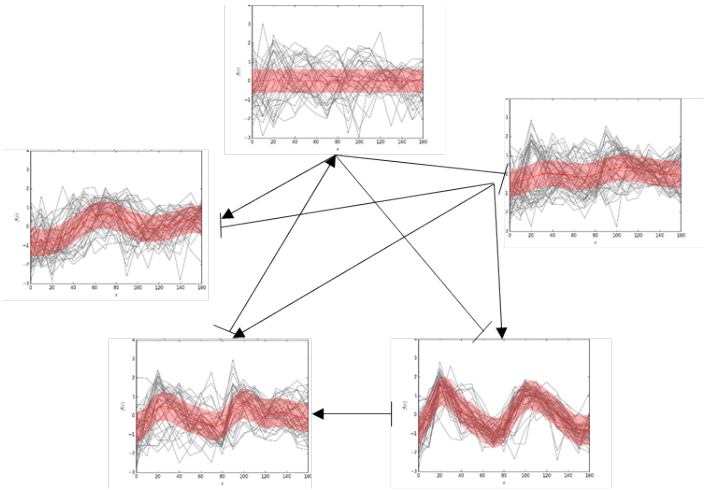


Figure 4.3: Cluster network inferred by ClusterNet [96]. Self-regulations are omitted.

Using the same gene expression data, a cluster network is generated by ClusterNet [96] with $k = 5$ for comparison with TiClNet. ClusterNet performs clustering itself and infer positive ($\rightarrow$) and negative ($\dashv$) relationships between clusters. As shown in Figure 4.3, some of the clusters show high variances among their gene expressions, making it hard to match the clusters with the cell cycle phase. Relatively low performance of ClusterNet is also measured by F1 score (0.280), adjusted rand index (0.081) and silhouette score (0.021). TiClNet inferred 8 edges among 5 clusters with positive and negative effects on each other but the relationship is complicated to explain without a specific order of the clusters.

4.4 Conclusion

In this study, TiClNet, a framework to improve clustering performance for time series is proposed. An adjusted version of gap statistics is implemented

in TiClNet to recommend the optimal number of clusters for the user-selected clustering algorithm. The adjusted gap statistics using Gaussian process is able to consider time-to-time dependency with the unevenly sampled time series, showing more accurate prediction of the number of clusters in cell cycle dataset. TiClNet provides a network of clusters where relationship between clusters are inferred by Shape-Based Distance and biological pathway enrichment test. With cell cycle dataset, TiClNet generated a correctly ordered network of clusters, each cluster matching to one of the cell cycle phases. Furthermore, a group of clusters with densely connected edges was identified as a functionally related module of DNA repair mechanism, showing the performance of TiClNet for biological interpretation.

Chapter 5

Conculsion

The increasing number of biological experiment measured at multiple time points requires novel methods and algorithms for analyzing such data. However, the analysis of gene expression time series is computationally challenging in that there are additional dimensions to consider such as condition (phenotype) and interaction between genes. This thesis presented three studies of network and clustering algorithms to solve the problem:

1. a network topology-based approach for pathway enrichment analysis that extends the existing method for time series and quantifies network propagation along time

2. a method to detect sub-network with network propagation along time and regulators of the sub-network using cross-correlation and influence maximization

3. a method to improve clustering by considering time dependency and construct the cluster network

In the first study, a network topology-based approach for pathway enrichment analysis, TRAP, is developed for analyzing time series transcriptome data. TRAP extends the existing pathway analysis method, SPIA, for time series analysis and estimates statistical values to measure the dynamic propagation of signaling effect in the pathway graph. In experiments on a proprietary dataset for the analysis of rice upon drought stress, TRAP was able to find relevant pathways more accurately than several existing methods. In the second study, a method to detect regulators of perturbed pathways, TimeTP, is developed. TimeTP performs pathway analysis first to determine a set of perturbed sub-pathways containing genes that are connected to propagate expression changes over time by measuring cross-correlation between two vectors of expression. To detect regulators of the perturbed pathways, TimeTP extends the gene network to include upstream regulators of genes such as transcription factors. Influence maximization technique is used to evaluate and rank the influence of regulators on the perturbed pathways. TimeTP was applied to PIK3CA knock-in dataset and found significant sub-pathways and their regulators relevant to the PIP3 signaling pathway. In the final study, a clustering method for gene expression time series is proposed. Although several clustering methods have been developed especially for time series data, most of the algorithms assume evenly spaced time series and do not take the time interval into account. However, biological experiments are often performed in unevenly spaced time intervals due to the experimental constraints on biological samples. This study aims to incorporate Gaussian process regression into the clustering process to predict unseen values of time series and provide more accurate clustering result. In addition, a network of clusters is generated by measuring distance (similarity) of expression patterns between clusters. As a distance measure, shape-based distance (SBD) is applied to capture similarity between time-shifted patterns.

This can infer gene regulatory relationship and cascading signaling effect over time. In conclusion, I have developed two algorithms for gene expression time series that quantifies network propagation along time to find perturbed pathways and detects sub-network (sub-pathway) with network propagation along time and regulators of the sub-network. The analysis results of two algorithms show that measuring and detecting network propagation is effective to find biological insight from gene expression time series. In addition, the algorithm I proposed in the last chapter is expected to improve clustering by time point estimation using Gaussian process regression and generate cluster network with similar expression pattern and biological function.

Bibliography

[1] C. M. O'Connor, J. U. Adams, and J. Fairman, "Essentials of cell biology," *Cambridge, MA: NPG Education*, vol. 1, 2010.

[2] Y. Zhang, E. Butelli, S. Alseekh, T. Tohge, G. Rallapalli, J. Luo, P. G. Kawar, L. Hill, A. Santino, A. R. Fernie, *et al.*, "Multi-level engineering facilitates the production of phenylpropanoid compounds in tomato," *Nature communications*, vol. 6, p. 8635, 2015.

[3] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS computational biology*, vol. 8, no. 2, p. e1002375, 2012.

[4] V. K. Ramanan, L. Shen, J. H. Moore, and A. J. Saykin, "Pathway analysis of genomic data: concepts, methods, and prospects for future development," *TRENDS in Genetics*, vol. 28, no. 7, pp. 323–332, 2012.

[5] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 299–313, 2014.

[6] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[7] D. Nam, J. Kim, S.-Y. Kim, and S. Kim, "Gsa-snp: a general approach for gene set analysis of polymorphisms," *Nucleic acids research*, p. gkq428, 2010.

[8] I. Medina, D. Montaner, N. Bonifaci, M. A. Pujana, J. Carbonell, J. Tarraga, F. Al-Shahrour, and J. Dopazo, "Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies," *Nucleic acids research*, vol. 37, no. suppl 2, pp. W340–W344, 2009.

[9] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[10] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[11] L. Jacob, P. Neuvial, and S. Dudoit, "Gains in power from structured two-sample tests of means on graphs," *arXiv preprint arXiv:1009.5173*, 2010.

[12] W. A. Haynes, R. Higdon, L. Stanberry, D. Collins, and E. Kolker, "Differential expression analysis for pathways," *PLoS Comput. Biol*, vol. 9, no. 3, p. e1002967, 2013.

[13] P. Martini, G. Sales, M. S. Massa, M. Chiogna, and C. Romualdi, "Along signal paths: an empirical gene set approach exploiting pathway topology," *Nucleic acids research*, vol. 41, no. 1, pp. e19–e19, 2013.

[14] T. Park, S.-G. Yi, S. Lee, S. Y. Lee, D.-H. Yoo, J.-I. Ahn, and Y.-S. Lee, "Statistical tests for identifying differentially expressed genes in time-course microarray experiments," *Bioinformatics*, vol. 19, no. 6, pp. 694–703, 2003.

[15] Z. Bar-Joseph, G. Gerber, I. Simon, D. K. Gifford, and T. S. Jaakkola, "Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes," *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10146–10151, 2003.

[16] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12837–12842, 2005.

[17] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talón, "masigpro: a method to identify significantly differential expression profiles in time-course microarray experiments," *Bioinformatics*, vol. 22, no. 9, pp. 1096–1102, 2006.

[18] T. Äijö, V. Butty, Z. Chen, V. Salo, S. Tripathi, C. B. Burge, R. Lahesmaa, and H. Lähdesmäki, "Methods for time series analysis of rna-seq data with application to human th17 cell differentiation," *Bioinformatics*, vol. 30, no. 12, pp. i113–i120, 2014.

[19] N. Leng, Y. Li, B. E. Mcintosh, B. K. Nguyen, B. Duffin, S. Tian, J. A. Thomson, C. Dewey, R. Stewart, and C. Kendziorski, "Ebseq-hmm: A

bayesian approach for identifying gene-expression changes in ordered rna-seq experiments," *Bioinformatics*, p. btv193, 2015.

[20] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi, "Large-scale temporal gene expression mapping of central nervous system development," *Proceedings of the National Academy of Sciences*, vol. 95, no. 1, pp. 334–339, 1998.

[21] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proceedings of the National Academy of Sciences*, vol. 99, no. 14, pp. 9121–9126, 2002.

[22] A. Schliep, A. Schönhuth, and C. Steinhoff, "Using hidden markov models to analyze gene expression time course data," *Bioinformatics*, vol. 19, no. suppl 1, pp. i255–i263, 2003.

[23] L. Zhao and M. J. Zaki, "Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 694–705, ACM, 2005.

[24] A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. Furlong, N. D. Lawrence, and M. Rattray, "Model-based method for transcription factor target identification with limited data," *Proceedings of the National Academy of Sciences*, vol. 107, no. 17, pp. 7793–7798, 2010.

[25] Y. Kim, J.-H. Jang, S. Choi, and D. Hwang, "Tempi: probabilistic modeling time-evolving differential ppi networks with multiple information," *Bioinformatics*, vol. 30, no. 17, pp. i453–i460, 2014.

[26] Y. Kim, S. Han, S. Choi, and D. Hwang, "Inference of dynamic networks using time-course data," *Briefings in bioinformatics*, vol. 15, no. 2, pp. 212–228, 2014.

[27] K. Jo, H.-B. Kwon, and S. Kim, "Time-series rna-seq analysis package (trap) and its application to the analysis of rice, oryza sativa l. ssp. japonica, upon drought stress," *Methods*, vol. 67, no. 3, pp. 364–372, 2014.

[28] P. Martini, G. Sales, E. Calura, S. Cagnin, M. Chiogna, and C. Romualdi, "timeclip: pathway analysis for time course data without replicates," *BMC bioinformatics*, vol. 15, no. Suppl 5, p. S3, 2014.

[29] J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph, "Reconstructing dynamic regulatory maps," *Molecular Systems Biology*, vol. 3, no. 1, p. 74, 2007.

[30] M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, *et al.*, "The transcriptional network for mesenchymal transformation of brain tumours," *Nature*, vol. 463, no. 7279, pp. 318–325, 2010.

[31] K. Jo, I. Jung, J. H. Moon, and S. Kim, "Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways," *Bioinformatics*, vol. 32, no. 12, pp. i128–i136, 2016.

[32] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature protocols*, vol. 4, no. 1, p. 44, 2008.

[33] Z. Bar-Joseph, A. Gitter, and I. Simon, "Studying and modelling dynamic biological processes using time-series gene expression data," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 552–564, 2012.

[34] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, *et al.*, "Ncbi geo: archive for functional genomics data sets—10 years on," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D1005–D1010, 2011.

[35] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences*, vol. 98, no. 9, pp. 5116–5121, 2001.

[36] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, Springer, 2005.

[37] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "Edge: extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, no. 4, pp. 507–508, 2006.

[38] M. Aryee, J. Gutierrez-Pabello, I. Kramnik, T. Maiti, and J. Quackenbush, "An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: Betr (bayesian estimation of temporal regulation)," *BMC bioinformatics*, vol. 10, no. 1, p. 409, 2009.

[39] J. Ernst and Z. Bar-Joseph, "Stem: a tool for the analysis of short time series gene expression data," *BMC bioinformatics*, vol. 7, no. 1, p. 191, 2006.

[40] P. Magni, F. Ferrazzi, L. Sacchi, and R. Bellazzi, "Timeclust: a clustering tool for gene expression time series," *Bioinformatics*, vol. 24, no. 3, pp. 430–432, 2008.

[41] J. Sivriver, N. Habib, and N. Friedman, "An integrative clustering and modeling algorithm for dynamical gene expression data," *Bioinformatics*, vol. 27, no. 13, pp. i392–i400, 2011.

[42] A. Sinha and M. Markatou, "A platform for processing expression of short time series (pests)," *BMC bioinformatics*, vol. 12, no. 1, p. 13, 2011.

[43] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[44] I. M. Ong, J. D. Glasner, and D. Page, "Modelling regulatory pathways in e. coli from time series expression profiles," *Bioinformatics*, vol. 18, no. suppl 1, pp. S241–S248, 2002.

[45] N. D. Mukhopadhyay and S. Chatterjee, "Causality and pathway search in microarray time series experiment," *Bioinformatics*, vol. 23, no. 4, pp. 442–449, 2007.

[46] L. Astola, M. Groenenboom, V. G. Roldan, F. Van Eeuwijk, R. D. Hall, A. Bovy, and J. Molenaar, "Metabolic pathway inference from time series data: a non iterative approach," in *Pattern Recognition in Bioinformatics*, pp. 97–108, Springer, 2011.

[47] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[48] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with rna-seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[49] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[50] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with rna-seq," *Nature biotechnology*, vol. 31, no. 1, p. 46, 2013.

[51] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Laboratory (LANL), 2008.

[52] H. Sakai, S. S. Lee, T. Tanaka, H. Numa, J. Kim, Y. Kawahara, H. Wakimoto, C.-c. Yang, M. Iwamoto, T. Abe, *et al.*, "Rice annotation project database (rap-db): an integrative and interactive database for rice genomics," *Plant and Cell Physiology*, vol. 54, no. 2, pp. e6–e6, 2013.

[53] S.-J. Oh, Y. S. Kim, C.-W. Kwon, H. K. Park, J. S. Jeong, and J.-K. Kim, "Overexpression of the transcription factor ap37 in rice improves grain yield under drought conditions," *Plant Physiology*, vol. 150, no. 3, pp. 1368–1379, 2009.

[54] N. Tuteja, "Abscisic acid and abiotic stress signaling," *Plant signaling & behavior*, vol. 2, no. 3, pp. 135–138, 2007.

[55] H. Cochard, L. Coll, X. Le Roux, and T. Améglio, "Unraveling the effects of plant hydraulics on stomatal closure during water stress in walnut," *Plant Physiology*, vol. 128, no. 1, pp. 282–290, 2002.

[56] P. C. Bethke, R. Schuurink, and R. L. Jones, "Hormonal signalling in cereal aleurone," *Journal of Experimental Botany*, vol. 48, no. 7, pp. 1337–1356, 1997.

[57] M. W. Yaish, A. El-Kereamy, T. Zhu, P. H. Beatty, A. G. Good, Y.-M. Bi, and S. J. Rothstein, "The apetala-2-like transcription factor osap2-39 controls key interactions between abscisic acid and gibberellin in rice," *PLoS genetics*, vol. 6, no. 9, p. e1001098, 2010.

[58] A. Repellin, A. Thi, A. Tashakorie, Y. Sahsah, C. Daniel, and Y. Zuily-Fodil, "Leaf membrane lipids and drought tolerance in young coconut palms (*Cocos nucifera* l.)," *European Journal of Agronomy*, vol. 6, no. 1, pp. 25–33, 1997.

[59] F. M. De Paula, A. Thi, J. V. De Silva, A. Justin, C. Demandre, and P. Mazliak, "Effects of water stress on the molecular species composition of polar lipids from *Vigna unguiculata* l. leaves," *Plant Science*, vol. 66, no. 2, pp. 185–193, 1990.

[60] M. Zhang, R. Barg, M. Yin, Y. Gueta-Dahan, A. Leikin-Frenkel, Y. Salts, S. Shabtai, and G. Ben-Hayyim, "Modulated fatty acid desaturation via overexpression of two distinct ω-3 desaturases differentially alters tolerance to various abiotic stresses in transgenic tobacco cells and plants," *The Plant Journal*, vol. 44, no. 3, pp. 361–371, 2005.

[61] J. M. McCord and I. Fridovich, "Superoxide dismutase an enzymic function for erythrocuprein (hemocuprein)," *Journal of Biological chemistry*, vol. 244, no. 22, pp. 6049–6055, 1969.

[62] F.-Z. Wang, Q.-B. Wang, S.-Y. Kwon, S.-S. Kwak, and W.-A. Su, "Enhanced drought tolerance of transgenic rice plants expressing a pea manganese superoxide dismutase," *Journal of plant physiology*, vol. 162, no. 4, pp. 465–472, 2005.

[63] Y. Tanaka, T. Hibino, Y. Hayashi, A. Tanaka, S. Kishitani, T. Takabe, and S. Yokota, "Salt tolerance of transgenic rice overexpressing yeast mitochondrial mn-sod in chloroplasts," *Plant Science*, vol. 148, no. 2, pp. 131–138, 1999.

[64] D. Spies and C. Ciaudo, "Dynamics in transcriptomics: Advancements in rna-seq time course and downstream analysis," *Computational and structural biotechnology journal*, vol. 13, pp. 469–477, 2015.

[65] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *Acoustics, speech and signal processing, IEEE transactions on*, vol. 30, no. 6, pp. 998–1003, 1982.

[66] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, ACM, 2003.

[67] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome Biol*, vol. 15, no. 12, p. 550, 2014.

[68] L. A. Bovolenta, M. L. Acencio, and N. Lemke, "Htridb: an open-access database for experimentally verified human transcriptional regulation interactions," *BMC genomics*, vol. 13, no. 1, p. 405, 2012.

[69] J. Ernst, H. L. Plasterer, I. Simon, and Z. Bar-Joseph, "Integrating multiple evidence sources to predict transcription factor binding in the human genome," *Genome research*, vol. 20, no. 4, pp. 526–536, 2010.

[70] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, *et al.*, "String v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic acids research*, p. gku1003, 2014.

[71] F.-H. Li, C.-T. Li, and M.-K. Shan, "Labeled influence maximization in social networks for target marketing," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 560–563, IEEE, 2011.

[72] N. E. Hynes and H. A. Lane, "Erbb receptors and cancer: the complexity of targeted inhibitors," *Nature Reviews Cancer*, vol. 5, no. 5, pp. 341–354, 2005.

[73] C. Jiménez, R. A. Portela, M. Mellado, J. M. Rodríguez-Frade, J. Collard, A. Serrano, C. Martínez-a, J. Avila, and A. C. Carrera, "Role of the pi3k regulatory subunit in the control of actin organization and cell migration," *The Journal of cell biology*, vol. 151, no. 2, pp. 249–262, 2000.

[74] F. Berglund, N. R. Weerasinghe, L. Davidson, J. C. Lim, B. J. Eickholt, and N. R. Leslie, "Disruption of epithelial architecture caused by loss of

pten or by oncogenic mutant p110α/pik3ca but not by her2 or mutant akt1," *Oncogene*, vol. 32, no. 37, pp. 4417–4426, 2013.

[75] M. A. Essers, L. M. de Vries-Smits, N. Barker, P. E. Polderman, B. M. Burgering, and H. C. Korswagen, "Functional interaction between ß-catenin and foxo in oxidative stress signaling," *Science*, vol. 308, no. 5725, pp. 1181–1184, 2005.

[76] J. M. Perry, X. C. He, R. Sugimura, J. C. Grindley, J. S. Haug, S. Ding, and L. Li, "Cooperation between both wnt/β-catenin and pten/pi3k/akt signaling promotes primitive hematopoietic stem cell self-renewal and expansion," *Genes & development*, vol. 25, no. 18, pp. 1928–1942, 2011.

[77] L. Vadlakonda, M. Pasupuleti, and R. Pallu, "Role of pi3k-akt-mtor and wnt signaling pathways in transition of g1-s phase of cell cycle in cancer cells," *Frontiers in oncology*, vol. 3, 2013.

[78] X. Zhang, N. Tang, T. J. Hadden, and A. K. Rishi, "Akt, foxo and regulation of apoptosis," *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1813, no. 11, pp. 1978–1986, 2011.

[79] Y. Zhang, A. D. Hoppe, and J. A. Swanson, "Coordination of fc receptor signaling regulates cellular commitment to phagocytosis," *Proceedings of the National Academy of Sciences*, vol. 107, no. 45, pp. 19332–19337, 2010.

[80] V. Y. Kiselev, V. Juvin, M. Malek, N. Luscombe, P. Hawkins, N. Le Novère, and L. Stephens, "Perturbations of pip3 signalling trigger a global remodelling of mrna landscape and reveal a transcriptional feedback loop," *Nucleic acids research*, vol. 43, no. 20, pp. 9663–9679, 2015.

[81] C. M. Coticchia, C. M. Revankar, T. B. Deb, R. B. Dickson, and M. D. Johnson, "Calmodulin modulates akt activity in human breast cancer cell lines," *Breast cancer research and treatment*, vol. 115, no. 3, pp. 545–560, 2009.

[82] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[83] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[84] J. Monks, D. Rosner, F. J. Geske, L. Lehman, L. Hanson, M. Neville, and V. Fadok, "Epithelial cells as phagocytes: apoptotic epithelial cells are engulfed by mammary alveolar epithelial cells and repress inflammatory mediator release," *Cell Death & Differentiation*, vol. 12, no. 2, pp. 107–114, 2005.

[85] M. Gilchrist, V. Thorsson, B. Li, A. G. Rust, M. Korb, K. Kennedy, T. Hai, H. Bolouri, and A. Aderem, "Systems biology approaches identify atf3 as a negative regulator of toll-like receptor 4," *Nature*, vol. 441, no. 7090, pp. 173–178, 2006.

[86] T. B. Deb, C. M. Coticchia, and R. B. Dickson, "Calmodulin-mediated activation of akt regulates survival of c-myc-overexpressing mouse mammary carcinoma cells," *Journal of Biological Chemistry*, vol. 279, no. 37, pp. 38903–38911, 2004.

[87] S. Lee, D. Kim, K. Lee, J. Choi, S. Kim, M. Jeon, S. Lim, D. Choi, S. Kim, A.-C. Tan, *et al.*, "Best: next-generation biomedical entity search tool for

knowledge discovery from biomedical literature," *PloS one*, vol. 11, no. 10, p. e0164680, 2016.

[88] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[89] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome biology*, vol. 3, no. 7, pp. research0036–1, 2002.

[90] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, *et al.*, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular cell*, vol. 2, no. 1, pp. 65–73, 1998.

[91] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1855–1870, ACM, 2015.

[92] E. J. Cooke, R. S. Savage, P. D. Kirk, R. Darkins, and D. L. Wild, "Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements," *BMC bioinformatics*, vol. 12, no. 1, p. 399, 2011.

[93] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[94] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[95] D. Defays, "An efficient algorithm for a complete link method," *The Computer Journal*, vol. 20, no. 4, pp. 364–366, 1977.

[96] Y. Shiraishi, S. Kimura, and M. Okada, "Inferring cluster-based networks from differently stimulated multiple time-course gene expression data," *Bioinformatics*, vol. 26, no. 8, pp. 1073–1081, 2010.

요약

차세대 염기서열 분석법의 대중화에 따라 전장 유전체 시퀀싱을 통한 유전자 발현 데이터 역시 활발하게 생산되고 있다. 특히 시계열 유전자 발현 데이터의 경우 특정 조건에서 일어나는 반응을 시간대 별로 측정하여 표현형에 따라 다르게 나타나는 생물학적 메커니즘을 보다 정확히 규명할 수 있다. 이에 따라 유전자 발현 데이터 가운데 시계열 데이터의 생산 비율이 점차 높아지고 있는 추세이나, 현재까지 개발된 대다수의 생물 데이터 분석 기법은 단일 시간대 데이터를 대상으로 하고 있다.

본 논문에서는 시계열 유전자 발현 데이터로부터 생물학적 패스웨이 및 클러스터의 네트워크를 구축하고, 해당 네트워크 상의 발현 패턴 전파를 분석하는 세 가지 생물정보학 기법을 소개한다. 첫 번째는 네트워크 위상을 활용한 시계열 유전자 발현 데이터의 생물학적 패스웨이 분석 기법(TRAP)이다. 해당 연구에서는 기존 패스웨이 분석 기법을 시계열 데이터에 맞게 확장하여 패스웨이 그래프에서 시간에 따른 신호 전달 여부를 수치화하고 통계량을 정의하였다. 개발된 기법으로 쌀 RNA 시퀀싱 데이터로부터 가뭄저항벼에서 활성화된 패스웨이를 선정한 결과, 아브시스산 합성, 지베렐린 합성 패스웨이와 같이 기존 기법으로 검출되지 않은 가뭄저항 관련 패스웨이가 확인되었다. 두 번째는 표현형에 따라 다르게 발현된 서브패스웨이 및 조절자를 찾는 기법(TimeTP)이다. 해당 기법은 유전자 간의 상호 상관(cross correlation) 수치를 통해 시간에 따라 발현이 전파된 유전자 쌍을 필터링하여 서브패스웨이를 검출하고, 전체 네트워크에 영향력 최대화(influence maximization) 기법을 적용하여 서브패스웨이에 포함된 유전자들을 조절하는 전사인자를 선정한다. 개발된 기법을 사용하여 PIK3CA 유전자 녹인(knock-in) 데이터를 분석한 결과, PIK3CA 유전자가 활성화시킨 PI3K 신호전달과 관련된 패스웨이들이 검출되었으며 기존 생물 연구에서 밝혀진 피드백 루프

또한 재현된 것을 확인할 수 있었다. 세번째는 시계열 유전자 발현 데이터로부터 클러스터 네트워크를 구축하는 기법(TiClNet)이다. 해당 기법은 가우시안 프로세스 회귀(Gaussian process regression)를 이용해 기존 클러스터링 기법에서 고려되지 않은 시간대 간 종속성을 고려할 뿐 아니라 및 비관측 시간대의 발현량을 추정하여 클러스터링을 정확성을 높이고자 하였다. 또한, 클러스터에 속한 유전자의 발현 패턴 및 생물학적 기능의 유사도를 계산하여 클러스터 간의 네트워크를 구축하였다. 벤치마크 데이터에 본 기법을 적용한 결과 실제 세포 주기와 순서 및 기능이 일치하는 네트워크를 생성하는 것을 확인하였으며, 기존 클러스터 네트워크 구축 기법에 비해 정확도 및 복잡도에서 더 뛰어난 결과를 보였다.

본 박사학위논문에서는 시계열 유전자 발현 데이터의 분석을 위해 네트워크 상에서 유전자 발현 패턴이 전파된 생물학적 패스웨이를 선정하는 기며, 서브패스웨이 및 조절 전사인자를 선정하는 기법, 클러스터 네트워크를 구축하는 기법을 개발하였다. 세 기법은 공통적으로 시계열 유전자 발현 데이터 분석의 난점인 표현형, 유전자, 시간 축의 통합 분석을 수행할 수 있으며, 각각의 성능 또한 기존 생물 데이터를 통해 증명되었으므로 추후 다양한 생물정보 분석에 활용될 것으로 기대된다.

주요어: 유전자 발현, 시계열, 생물학적 패스웨이, 네트워크, 클러스터링
학번: 2013-20884