공학박사 학위논문

# Machine Learning Approaches for Decoding Information in RNA Interactions and DNA sequences

## RNA 상호작용 및 DNA 서열의 정보해독을 위한 기계학습 기법

2020 년 2 월

서울대학교 대학원

컴퓨터 공학부

이 상 선

공학박사 학위논문

# Machine Learning Approaches for Decoding Information in RNA Interactions and DNA sequences

## RNA 상호작용 및 DNA 서열의 정보해독을 위한 기계학습 기법

2020 년 2 월

서울대학교 대학원

컴퓨터 공학부

이 상 선

# Machine Learning Approaches for Decoding Information in RNA Interactions and DNA sequences

## RNA 상호작용 및 DNA 서열의 정보해독을 위한 기계학습 기법

지도교수 김 선

이 논문을 공학박사 학위논문으로 제출함

2019 년 11 월


서울대학교 대학원

컴퓨터 공학부

이 상 선


이상선의 공학박사 학위논문을 인준함

2019 년 12 월

| | |
|---|---|
| 위 원 장 | 김형주 |
| 부위원장 | 김선 |
| 위 원 | 강유 |
| 위 원 | 이상혁 |
| 위 원 | 채희준 |

# Abstract

## Machine Learning Approaches for Decoding Information in RNA Interactions and DNA sequences

Sangseon Lee
Department of Computer Science & Engineering
College of Engineering
Seoul National University

Phenotypic differences among organisms are mainly due to the difference in genetic information. As a result of genetic information modification, an organism may evolve into a different species and patients with the same disease may have different prognosis. This important biological information can be observed in the form of various omics data using high throughput instrument technologies such as sequencing instruments. However, interpretation of such omics data is challenging since omics data is with very high dimensions but with relatively small number of samples. Typically, the number of dimensions is higher than the number of samples, which makes the interpretation of omics data one of the most challenging machine learning problems.

My doctoral study aims to develop new bioinformatics methods for decoding information in these high dimensional data by utilizing machine learning algorithms.

The first study is to analyze the difference in the amount of information between different regions of the DNA sequence. To achieve the goal, a ranked-based $k$-spectrum string kernel, RKSS kernel, is developed for comparative and evolutionary comparison of various genomic region sequences among multiple species. RKSS kernel extends the existing $k$-spectrum string kernel by utilizing rank information of $k$-mers and landmarks of $k$-mers that represents a species. By using a landmark as a reference point for comparison, the number of $k$-mers needed to calculating sequence similarities is dramatically reduced. In the experiments on three different genomic regions, RKSS kernel captured more reliable distances between species according to genetic information contents of the target region. Also, RKSS kernel was able to rearrange each region to match a biological common insight.

The second study aims to efficiently decode complex genetic interactions using biological networks and, then, to classify cancer subtypes by interpreting biological functions. To achieve the goal, a pathway-based deep learning model using graph convolutional network and multi-attention based ensemble (GCN+MAE) for cancer subtype classification is developed. In order to efficiently reduce the relationships between genes using pathway information, GCN+MAE is designed as an explainable deep learning structure using graph convolutional network and attention mechanism. Extracted pathway-level information of cancer subtypes is transported into gene-level again by network propagation. In the experiments of five cancer data sets, GCN+MAE showed better cancer subtype classification performances and captured subtype-specific pathways and their biological functions.

The third study is to identify sub-networks of a biological pathway. The goal is to dissect a biological pathway into multiple sub-networks, each of which is to be of a single functional unit. To achieve the goal, a condition-specific sub-module detection method in a biological network, MIDAS (MIning Differentially Activated Subpaths) is developed. From the pathway, edge activities are measured by explicit gene expression and network topology. Using the activities, differentially activated subpaths are explored by a statistical approach. Also, by extending this idea on graph convolutional network, different sub-networks are highlighted by attention mechanisms. In the experiment with breast cancer data, MIDAS and the deep learning model successfully decomposed gene-level features into sub-modules of single functions.

In summary, my doctoral study proposes new computational methods to compare genomic DNA sequences as information contents, to model pathway-based cancer subtype classifications and regulations, and to identify condition-specific sub-modules among multiple cancer subtypes.

**Keywords**: High dimensional data, Biological prior knowledge, DNA sequence, Gene expression, Machine learning

**Student Number**: 2014-21754

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Genetic information accumulates as organisms evolve. Interpreting the genetic information is very important to help reveal the secrets of living things. Thanks to advances in instrument technologies, genetic information has increased dramatically. Sequences of genomes, or DNA sequences, of many species are now available and it is possible to compare genomes of species by comparing genome sequences. In addition, RNA-sequencing technologies produced condition-specific gene expression profiles. Interpreting gene expressions and interactions can clarify the causes of external, physical, and pathological differences among people. In my doctoral study, I developed machine learning algorithms and methods to compare and interpret DNA sequences and gene expression profiles. I used DNA sequence information to compare different species and gene expression information to compare and stratify cancer patients in the form of cancer subtypes.

## 1.1 Biological questions with genetic information

### 1.1.1 Biological Sequences

A genome is a DNA sequence that contains all the genetic information of an organism. Depending on genetic information in the genome, phenotypes of organisms can be different and an organism may evolve to a new species. Thus, interpreting the information from genome sequences can help understand differences among species. However, decoding genomes is challenging due to the huge size of a genome, 3.2 billion nucleotides in the human genome. A genome is a sequence of nucleotides (Adenine, Cytosine, Guanine, and Thymine) without grammatical structure such as words and sentences. However, a genome consists of distinct structural components. A genome can be divided into genes and non-genetic parts (Figure 1.1). A gene in the eukaryotic genome consists of multiple components or subsequences such as exons that contain genetic information and introns between exons. Additionally, non-genetic regions can be divided into multiple components such as CpG islands, promoters, and enhancers. These non-genetics regions are known to be involved in regulating expression of genes. Therefore, identifying differences among these genomics regions is an essential task to interpret the genetic information of genome sequences.

### 1.1.2 Gene expression

Gene expression information in a cell can represent activities of biological functions in an organism. Depending on which functions are turned on, phenotypic differences, such as appearance and disease, are determined. Thus, interpretation of gene expression data can be a clue to elucidate the unknown factors of why people are different and suffer from lethal diseases. The main challenge in analyzing gene expression data, however, is that genes perform

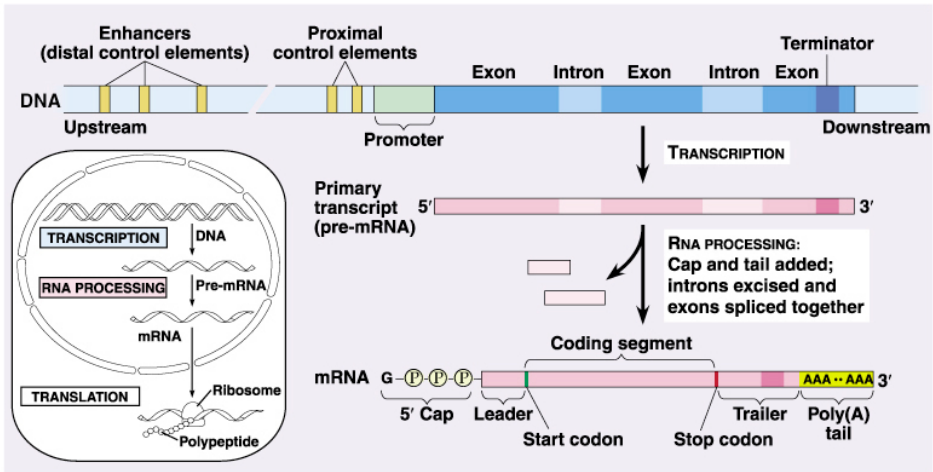



**Figure 1.1:** Genome structure composed of various regions such as exon, intron, or promoters.

biological functions through complex interactions among genes (Figure 1.2). Understanding the biological phenomena is to solve puzzles of the complicated genetic interactions. For example, abnormal expression of certain genes in a tumor can result in unusual aggressive growth of tumor. Therefore, decoding interactions of genes is an essential step to figure out the reason why organisms are different biologically.

## 1.2 Formulating computational problems for the biological questions

### 1.2.1 Decoding biological sequences by $k$-mer vectors

Different regions of a genome have different biological functions due to difference in genetic information in DNA sequences. A common approach to interpreting the amount of information contained in DNA sequences is to utilize

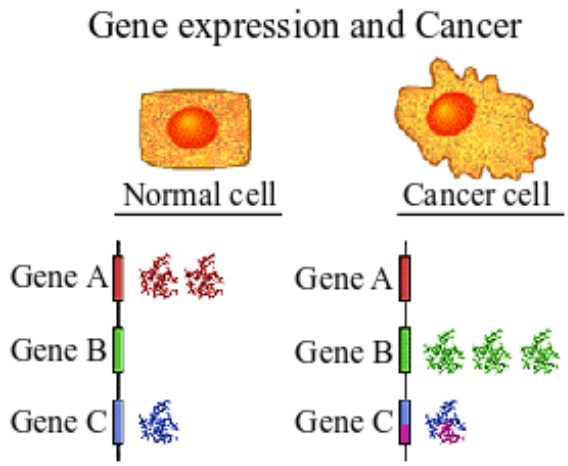



**Figure 1.2:** Regulation of biological phenomena through collaboration of multiple genes (from the Cancer Genome Anatomy Project (CGAP), Conceptual Tour, July 21, 2000.).

information of DNA composition of the sequence. However, unlike a general text sentence represented by a combination of words separated by spaces, a DNA sequence is simply a continuous sequence of characters without any visual structural components. Thus, decoding DNA sequence is a very challenging task. Rather than decoding DNA sequences at the character level, $k$-mer based methods have been developed for years. By measuring the frequency of each $k$-mer while scanning the sequence into a set of overlapping substring ($k$-mer) of length $k$, it is possible to extract characteristics of the sequence and express the sequence in the form of a vector. Since the original sequence is divided into $k$-mers, information of the sequence can be lost to some extent, but by converting an encoded sequence into a vector, many existing computational methods can be used.

However, it is still difficult to interpret information contained in DNA sequence using $k$-mer vector. First of all, $k$-mer vector is high-dimensional

data. Since there are four types of DNA bases, a total of $4^k$ $k$-mers can be detected along length of k. That is, the number of $k$-mers that can appear in the sequence increases exponentially with length of k. If one sequence is analyzed, $k$-mers that do not appear in the sequence may be excluded from the analysis, but when analyzing multiple sequences at the same time, many kinds of $k$-mers must be analyzed.

Moreover, the use of $k$-mer vectors is not easy. Figure 1.3 shows the result of inferring the amount of information as Shannon entropy using the $k$-mer vector of sequences belonging to the chimpanzee's exon, intron and CpG islands. Given that the larger Shannon entropy is, the smaller the amount of genetic information is encoded, the Shannon entropy magnitude relationship between exon and intron is consistent with biological knowledge and with previous entropy-based studies. However, the fact that the CpG island region has the highest amount of genetic information (smallest Shannon entropy) is not consistent with biological knowledge. Therefore, Shannon entropy is not an appropriate method for interpreting the amount of information, and there is a need for another method that can more accurately infer the difference in the amount of genetic information of various parts of the genome.

The problem of interpreting genetic information from DNA sequences is the method of inferring the amount of information using the characteristics of a $k$-mer vector of one sequence, such as Shannon entropy, and measuring the distance between two sequences to determine the amount of information between sequences. The general formulation of the two methods is as follows.

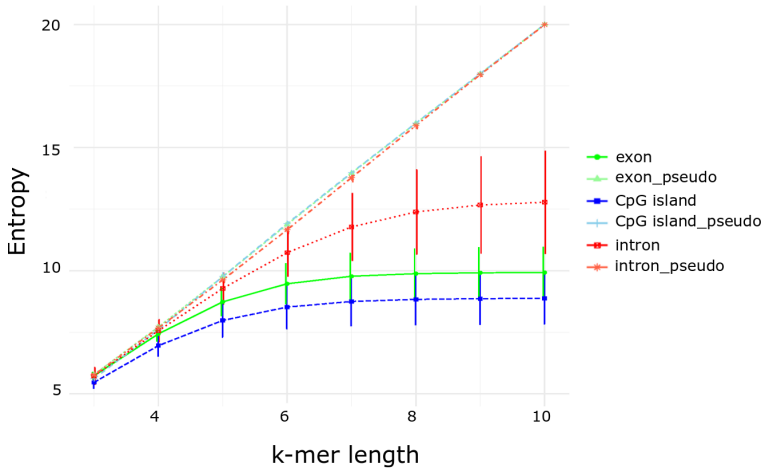



**Figure 1.3:** Shannon Entropy of three different region sequences with different length of $k$-mer (Species: Chimpanzee) .

$< Input >$

$x$ : a finite length of sequence with alphabet $\mathcal{A}$

$\mathcal{A} : \{A, C, G, T\}$

$< Output : UNKNOWN >$

$I(x)$ : an information of sequence $x$ OR

$D(x, x')$ : a distance between sequence $x$ and $x'$

$< Model >$

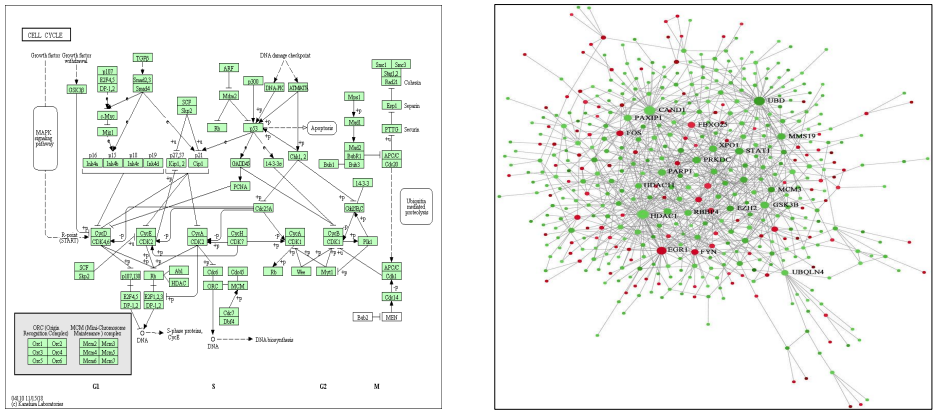
Transform a $x$ into $k$-mer vector space $\mathbb{R}^{4^k}$,

Define a metric $I$ to measure an information of $x$ OR

Define a distance measurement $D$ to distinguish sequence $x$ and $x'$

### 1.2.2 Interpretation of complex relationships between genes

Genes play function through complex interaction, thus a graph is widely used to model gene interactions. In a graph $G = (V, E)$, $V$ is a set of genes, each node in $V$ has a real number representing gene expression quantity, and $E$ is a set of edges that represent interaction between genes. The advantage of using a graph for the interpretation of gene interaction is that valuable biological knowledge can be easily embedded into graphs. An example of a biological network is the pathway database. KEGG pathway database (Kanehisa and Goto, 2000) is the most widely used a graph database that consists of hundred graphs, each of which represent well curated biological process. Another example is protein-protein interaction (PPI) network where known gene interactions are modeled as edges between two genes or proteins (Figure 1.4). Thus, there have been numerous studies that use biological networks or graphs. However, the main challenge in analyzing gene interaction graphs is the size of graphs. A graph contains as many nodes as 20,000 genes and the number of edges is typically over 100,000. Since a single patient is represented as a graph, analysis of gene expression data from patients requires to analyze a set of big graphs that correspond to the number of patients, typically hundreds to thousands. Thus, interpretation of gene interactions under specific conditions, e.g., cancer, is a problem of mining big graph data. The problem can be formulated as a problem of classifying labels of graphs having the same topology but different node values. In this case, in order to predict the label of the graph, it is necessary to be able to extract the interaction between the specific genes present on the graph, that is, the biological function. Then, a model can be generated to classify cancer subtypes by using the extracted gene interaction as a feature. The general formulation of this method is as follows.




(a) Example of Pathway

(b) Example of PPI

**Figure 1.4:** Example of biological network for representing gene interactions.

$< Input >$

$G:$ a graph (pathway or PPI network), $G = (V, E)$

$V:$ a set of genes, $|V|$ = Number of genes $\simeq 20,000$

$E:$ a set of gene interactions

$X:$ Input matrix of gene expression, $X \in \mathbb{R}^{N \times |V|}$

$< Output >$

$Y:$ Cancer subtype of given N samples

$Y = \{0, 1, 2, ..., c\}^N, c$ : number of classes

$< Model >$

Given a graph $G$ and a input $X$,

Learn a model $M$ to utilize high level information of pathways s.t

$M(X) = X' \in \mathbb{R}^{N \times \{m_1(V), m_2(V), ..., m_k(V)\}}$

$m_k(V)$ : $k$th gene interactions captured by model $M$ and graph $G$

Build a classifier $f$ to predict labels, $f(M(X)) = Y'$

## 1.3 Three computational problems for the biological questions

The common issue of the three biological questions in my doctoral research that biological omics data such as DNA sequences and gene expression are high dimensional data but of small number of samples. $K$-mers of a biological sequence increase exponentially with the length of $k$, and features of gene expression data increase exponentially with the interactions between genes. Therefore, in order to address the biological questions by using the computer approach, it is necessary to effectively reduce the features such as reducing the number of $k$-mers or removing unnecessary gene interactions through the biological network.

In order to analyze such high dimensional biological data on comparison of species or patients, my doctoral study defines one problem related to DNA sequences and two problems related to gene expression analysis. 1) The problem of inferring regional information differences in the genome using evolutionary similarities of different species. 2) The problem of classifying cancer subtypes by extracting and combining information from multiple biological networks. 3) The problem of classifying cancer subtypes by extracting subgraphs from biological networks. Detailed problems and solutions for each problem are as follows.

- **Problem 1) Ranked $k$-spectrum kernel for comparative and evolutionary comparison of DNA sequences (RKSS kernel) (Lee *et al.*, 2019b)**:

  **Challenges:** As part of the study of interpreting the information encoded in the DNA sequence, studies have been conducted to determine the distance between sequences based on the similarity of the sequences.

Previous studies have proposed a string kernel based on $k$-mer vectors and have been successfully used to compare biological sequences (Leslie *et al.*, 2001; Cuturi and Vert, 2005; Murray *et al.*, 2017). However, when the comparison extended to multiple genomes, the methods showed limitations. The genome has several regions with different patterns of nucleotide sequences, and the lengths of the sequences belonging to the regions vary widely. There is also a case where a substring of a certain length is repeated. Conventional $k$-mer vector-based string kernels are vulnerable to sequences of various lengths and repeated substrings because they use all $k$-mers of length k as well as their actual occurrence as feature values.

**Approach:** The Ranked k-spectrum string (RKSS) kernel addresses this problem with two ideas. 1) To reduce the effects of repeated substrings, the RKSS kernel uses $k$-mer's rank information instead of the actual frequency. 2) Assuming that several species evolved in differentiation from one common ancestor, the RKSS kernel defines a set of $k$-mers called landmark and uses it to compare species. Landmark is the highest $k$-mers commonly detected in various species. It not only reduces the high dimension of $k$-mer but also plays a role as a virtual common ancestor, and is a reference point for comparison between species. Based on these features, the RKSS kernel calculates the similarity between two sequences and uses it to determine the distance between sequences. For the 10 mammalian genomes and three regions (exon, intron, CpG islands), the RKSS kernel reproduces the phylogenetic tree more accurately than the existing string kernel method. In addition, a space called landmark space is defined by using several landmarks, and the order of information contents between three regions within the space is measured

in accordance with existing biological knowledge.

- **Problem 2) Pathway-based cancer subtype classification and interpretation by attention mechanism and network propagation (GCN+MAE) (Lee *et al.*, 2019a)**:

  **Challenges:** Pathway is a biological network that organizes the interactions between genes in graph form and is a small graph that contains only some genes that belong to a specific biological mechanism. Pathways can be used to efficiently analyze complex gene interactions and to easily interpret results. However, existing pathways contain only a few of the genes, one-third of human genes, resulting in unintended loss of information. Conventional pathway analysis tools generate a single value called pathway activity from gene expression levels, and models using these values calculated from several pathways have shown poor performance in the classification of cancer subtypes.

  **Approach:** In order to effectively predict cancer subtypes using pathway information, an interpretable deep learning model is proposed by using graph convolutional networks and multiple attention mechanisms. A graph convolutional network is used to capture gene patterns specifically expressed for each pathway and to generate pathway information vectors. The outputs of several graph convolution models are combined into two levels of attention layers. As a result, it is possible to extract pathways that are significant in predicting cancer subtypes. In addition, the Pathway-PPI network is constructed to compensate for the missing genes while simultaneously finding transcription factors that may contribute to the regulation of the pathways. By analyzing this through a network propagation algorithm, it is possible to detect subtype-specific

transcription factors and regulatory mechanisms. The model shows better performance than previous methods using pathway activity in predicting cancer subtypes in five real cancer data.

- **Problem 3) Detecting sub-modules in biological networks with gene expression by statistical approach and graph convolutional network (MIDAS) (Lee *et al.*, 2017)**:

  **Challenges:** Although pathway is a small graph of genes involved in similar biological phenomena, pathway does not perform a single biological function. An apoptosis pathway, for example, consists of genes that promote cell death and genes that inhibit cell death. In this way, various biological functions exist in one pathway, and the interaction between genes involved in the same biological functions within the pathway is called subpath. For this reason, the trend of pathway-based studies is gradually shifting towards finding subpaths. However, the existing methods use the activity of gene interactions with statistical values, not actual expression levels (Martini *et al.*, 2012; Nam *et al.*, 2014). There is also a limitation in discovering only subpaths useful for distinguishing two classes.

  **Approach:** MIDAS features two ways to find subpaths from a pathway. 1) It calculates the activity value between genes using the actual gene expression and pathway topology. 2) It can be applied to three or more class data by using statistical techniques. After finding a gene pair with high activity, the subpath is expanded by greedy expansion. The extension of the subpath iterates until the classification score is lower than the threshold, and the criterion is set to exponentially decaying as the iteration progresses. The application of MIDAS to breast cancer data has

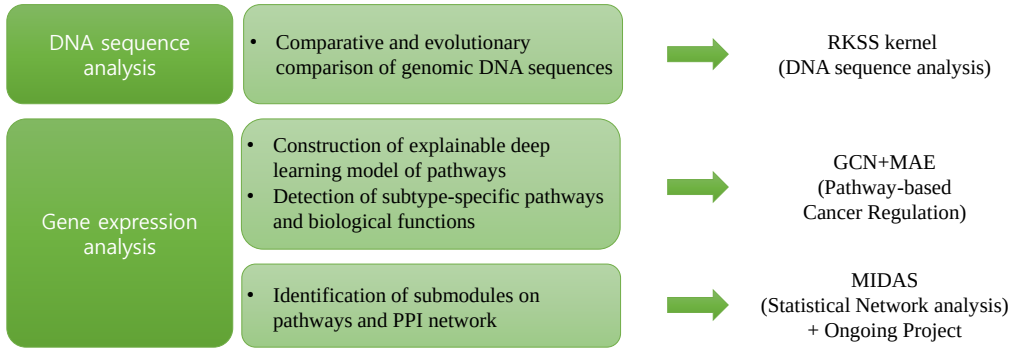



**Figure 1.5:** Computational challenges and solutions in DNA sequences and gene expression data.

shown better performance in predicting cancer subtypes and predicting patient survival than PATOME, a conventional subpath detection tool. Since the usefulness of extracting subpaths from the pathway is checked, extension MIDAS to a PPI network is performing as further study. The further trial with graph convolutional networks and class activation maps shows reasonably good performance in predicting cancer subtypes. In addition, subtype-specific subnetworks are extracted.

## 1.4 Outline of the thesis

Chapters 2,3, and 4 introduce independent studies related to machine learning algorithms of high dimensional data analysis for DNA sequence and gene expression data. In Chapter 2, a ranked k-spectrum kernel, RKSS kernel aims to successfully compare genomic sequences such as exon, intron, CpG island on 10 mammalian species. Chapter 3 describes an explainable deep learning method for pathway-based cancer subtype classification on five TCGA cancer datasets. Chapter 4 proposes a method for identifying submodules in biological networks; a pathway-based method named MIDAS, and PPI network based graph convolutional networks with attention mechanisms.

Chapter 5 summarizes the studies with my contributions in biological sequence and pathway-based gene expression based analyses. The thesis is concluded by an appendix of the bibliography of the cited references.

# Chapter 2

# Ranked k-spectrum kernel for comparative and evolutionary comparison of DNA sequences

A genome consists of distinct regions such as exon, intron, and CpG island, and each region has different biological functions. One way to interpret this genomic information is to measure the information contents of regions. Among the various methods for measuring information, an efficient way can be used for various regions of the genome sequence is to extract the characteristics of sequences based on $k$-mers. This study transforms a genomic sequence into a fix-length $k$-mer representation vector and devises a new computational method that focuses on the characteristics of genomic sequences and comparisons between different species. Based on two experiments using 10 mammalian species with exon, intron, and CpG island sequences, this analysis suggests that the relational order, exon > CpG island > intron, in terms of evolutionary information contents.

## 2.1 Motivation

Biological molecules in the cell such as DNA or proteins are commonly represented as sequences. For this reason, the biological functions of DNA or proteins have been investigated by comparing and characterizing biological sequences. Thus, biological sequence analysis has been at the heart of bioinformatics research (Durbin *et al.*, 1998). Due to the recent development of high throughput sequencing techniques and accordingly the increasing number of genome sequencing projects, sequence analysis methods have become even more important, and they have been extensively used for investigating important research topics.

Smith-Waterman (Smith and BEYER, 1976) and BLAST (Altschul *et al.*, 1990; Gish and States, 1993) have been successfully used to compute the similarity of sequences for a long time since the methods were introduced to the biology community. Then, a number of alignment-based methods have been developed for investigating important questions in biology. For detecting distant homologous relationships between proteins, the multiple sequence alignment of protein sequences was utilized with a profile hidden Markov model (profile HMM) (Söding, 2004). For prediction of transcription-factor target sites in the promoter regions, the phylogenetic footprinting approach used the information of orthologous sequences (Berezikov *et al.*, 2004). In addition, in an attempt to investigate evolutionary processes, there have been comparative researches based on RNA sequencing of multiple species (Perry *et al.*, 2012).

Alignment-based methods, although successfully used in many applications, are not computationally efficient to handle a large number of sequences that are generated by high throughput sequencing technologies. In addition, the methods provide information encoded in a sequence only in terms of alignment, in other words, information relative to a reference sequence. Thus,

there have been efforts to develop alignment-free based methods (Vinga and Almeida, 2003; Vinga, 2007, 2014; Bonham-Carter *et al.*, 2013). Basically, these methods are based on the frequency vector of $k$-length contiguous substrings called as $k$-mers. Once the $k$-mer vectors or discrete distributions are obtained from sequences, a similarity between the two sequences is measured in various ways like Euclidean distance methods (Blaisdell, 1986), cosine similarity (Stuart *et al.*, 2002), Kullback-Leibler discrepancy (Wu *et al.*, 2001; Das *et al.*, 2018), and methods of revising distance from evolutionary models (Allman *et al.*, 2017).

### 2.1.1 String kernel for sequence comparison

String kernel-based methods were originally proposed for classification of text documents using support vector machines (SVMs) (Watkins, 1999; Haussler, 1999; Lodhi *et al.*, 2002). When input data are strings, and I have their representations on a Euclidean space $\mathbb{R}^d$, I can calculate the string similarities using kernel functions and obtain the distance information in the space associated with the kernel. When input data are string and can be represented on a Euclidean space $\mathbb{R}^d$, string similarities and distances can be measured by using string kernel functions. For the biological sequence analysis, one of the earliest applications of the string kernel was the $k$-spectrum kernel for the protein sequence classification using SVM (Leslie *et al.*, 2001). In that study, protein sequences were projected into a $k$-mer feature space and similarity was measured by the inner product in that space.

Since then, various string kernels using a $k$-mer frequency vectors have been developed. To reflect the fact that biological sequences of same functionalities can be altered over time, resulting in substitutions, deletions, and insertions, $m$-mismatch and $k$-spectrum kernels using mismatch trees were developed (Leslie *et al.*, 2004). In a similar manner, weighted degree kernels

were designed, summing up string kernels with different $k$-mers (Smola and Vishwanathan, 2003; Rätsch *et al.*, 2005; Ben-Hur *et al.*, 2008) because the similarity measure may be affected by the value of $k$. Meanwhile, there are string kernel methods that utilize other data structures. For example, the hash table and Shannon entropy were used for the weighted sum of hashed $k$-mers (Murray *et al.*, 2017). Various statistical/evolutionary background models also adapted to measure sequence similarities, including D2static (Forêt *et al.*, 2009; Song *et al.*, 2013), Jukes and Cantor 1969 model (JC69) (Allman *et al.*, 2017), and a scoring matrix such as BLOSUM (Nojoomi and Koehl, 2017a,b). Accordingly, string kernels have been actively studied, and the information that the aforementioned methods have commonly used is the $k$-mer frequency vector to obtain the string similarities.

A different kernel-based approach to sequence comparison is to utilize an implicit representation of the sequence. As an example, the mutual information kernel (Seeger, 2002) measured the similarity of two sequences with probabilistic models, which need a strong assumption on the prior in the model. This type of kernels were implemented by various methods such as Markov chain process based context-tree model (Cuturi and Vert, 2005), profile HMM (Fong *et al.*, 2014), and Kullback-Leibler relative entropy (Ulitsky *et al.*, 2006). Another example of the string kernel using implicit representation is the alignment kernel. To mimic the score of Smith-Waterman algorithm (Smith and BEYER, 1976) when comparing two sequences, a local alignment kernel was designed with the appropriate mathematical basis (Saigo *et al.*, 2004). This local alignment string kernel was expanded by considering all possible alignments of $k$-mers with ignoring gaps (Shen *et al.*, 2014).

### 2.1.2 Approach: RKSS kernel

Many string kernel methods have been developed since Leslie's $k$-spectrum string kernel. However, existing kernel methods have limitations in explanatory power for comparative and evolutionary comparison of multiple species. To perform the comparative and evolutionary study of multiple genomes, the $k$-spectrum kernel produces a pairwise distance of two genomes. Combining many pairwise distances is not straightforward. More seriously, the $k$-spectrum kernel is sensitive to over-represented $k$-mers from repeats or gene duplications as shown in Section 2.2.2. Meanwhile, the alignment-based kernels require the sequence alignment information, and they have two serious limitations. First, obtaining alignment information requires a huge amount of computation time for a large number of sequences. Second, the alignment information is relative to each other and combining numerous alignments of a large number of sequences is a very complicated task. Therefore, new string kernel method is needed for comparative and evolutionary comparison of multiple species.

In this study, a novel *ranked k-spectrum string (RKSS) kernel* is proposed that can be used to construct phylogenetic trees and perform the comparison of exon, intron, and CpG island sequences. The basic idea is to select $k$-mer strings with respect to (a) reference point(s), or landmark(s). For the phylogenetic construction, a single landmark of $k$-mers that are common to the genomes is utilized in comparison. Then, a distance between two genomes is defined by proposed kernel method for comparing two constructed $k$-mer vectors according to a single landmark. For comparison of exons, intron, and CpG island sequences, three landmarks for exons, intron, and CpG island sequences are created on each species and distances between a pair of sequences are defined in terms of distance to the landmarks of all species.

In the literature, the reference or landmark-based analysis has been used for a number of sequence analysis tasks. Chae *et. al.* (Chae *et al.*, 2013) used

a set of common $k$-mer strings in CpG island sequences to construct phylogenetic trees of 10 mammalian genomes and performed machine learning analysis. Middleton *et. al.* (Middleton and Kim, 2014) utilized 1,973 RNA family covariance models from the Rfam database (Burge *et al.*, 2012) to define a new distance metric between RNA sequences. Using this distance metric, RNA structure motifs were identified without an additional process for sequences like alignment or folding. More recently, a $k$-mer based clustering method (Steinegger and Söding, 2018) that used reference sequences for defining and merging clusters has been proposed. My contribution in this study is to define *ranked $k$-spectrum string kernel*, *RKSS kernel*, for comparative and evolutionary sequence comparison using landmark (or reference) set of $k$-mer strings.

**[Problem Definition of this study]**

Given genomic regions $RG = \{\text{exon, intron, CpG island}\}$ and 10 mammalian species $MS, |MS| = 10$

$< Input >$

$X_{rg}^{ms}$ : a set of finite length sequences belonging to $rg \in RG$ and $ms \in MS$

$\mathcal{A} : \{A, C, G, T\}$, character set of sequences

$< Output >$

Relative information content of $RG$ when comparing $MS$

$< Model >$

Transform $X_{rg}^{ms}$ into a $k$-mer vector space $\mathbb{R}^{4^k}$,

Define a kernel $K$ to measure similarities between

two sequences $x$ and $x'$

Using the kernel $K$, compute similarities among $MS$

on a specific region $rg$

Based the similarities of $MS$ on each region $rg$,

measure relative information contents of $RG$ in evolutionary context

## 2.2 Methods

In this section, proposed methods for construction of the RKSS kernel and feature spaces are described. Also, a workflow is explained for applications of the RKSS kernel to the comparative and evolutionary analysis of 10 mammalian genomes: constructing phylogenetic trees and comparison of exons, introns, and CpG islands. The overview of this study is illustrated in Figure 2.1.

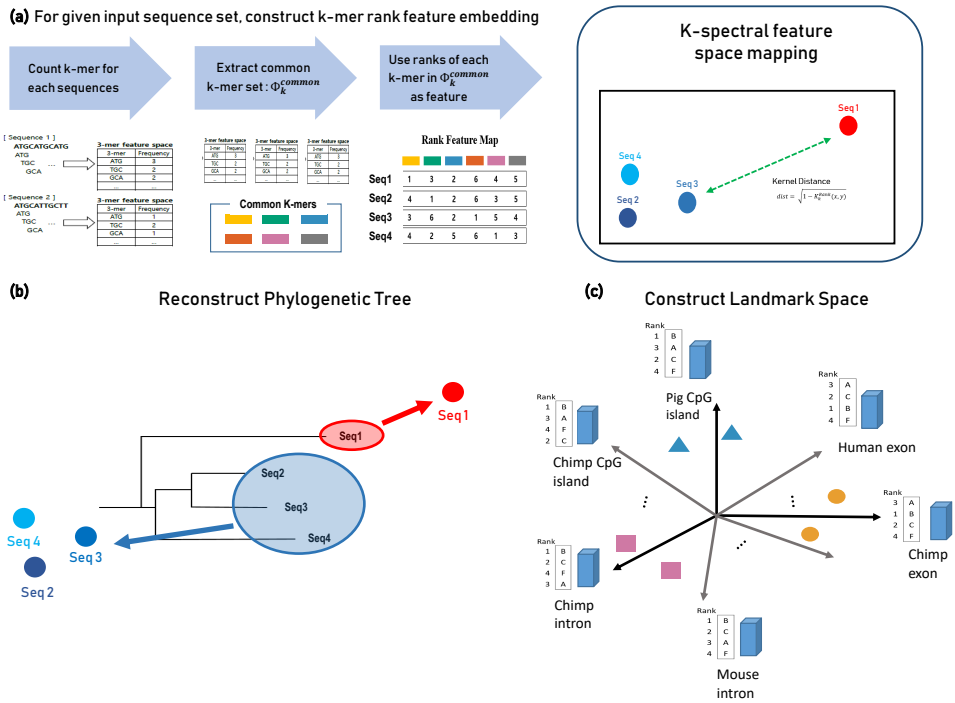



**Figure 2.1:** The workflow of ranked $k$-spectrum string kernel approach. (a) The step-wise calculation of ranked $k$-spectrum string kernel. For given input sequence set, $k$-mer frequencies are counted on each sequence. From that, a common $k$-mer template is extracted and used for the construction of rank feature map. Using the rank feature map, each sequence is mapped into $k$-spectral feature space and similarity between two points are measured by the kernel distance metric. (b) Example of reconstructing phylogenetic tree using a single landmark. A single $k$-mer template, a landmark, is used to measure pair-wise distances between species. (c) Example of using multiple landmarks. To distinguish sequences belonging to three genomic regions (exon, intron, CpG island), multiple $k$-mer sets with the rank profiles (multiple landmarks) construct a new feature space. Each input sequence is mapped into the space and similarity between sequences is computed according to their respective distances to all landmarks.

### 2.2.1 Mapping biological sequences to $k$-mer space: the $k$-spectrum string kernel

Before introducing a ranked string kernel for comparative and evolutionary sequence comparison, to help the reader understand, I explain what a string kernel is and how it defines the similarity and distance between two sequences. The kernel method is a similarity function over a pair of data points in the input space that implicitly transforms data into a new feature space and computes the inner products in that space. This approach also called *kernel trick*, is often used to classify non-linear data as linear discriminators (ex. Gaussian kernel).

The simplest and most successful way of constructing feature space for the biological sequences is the use of a set of $k$-length contiguous subsequences called $k$-mers. It is a concept similar to the bag-of-words model in natural language processing. On the input space $\mathcal{X}$ of all finite length sequences of characters from an alphabet $\mathcal{A}$, $|\mathcal{A}| = l$ ($l = 4$ for DNA sequences), a feature map $\Phi_k$ from $\mathcal{X}$ to $\mathbb{R}^{l^k}$ is defined as (Leslie *et al.*, 2001):

$$\Phi_k(x) = (\phi_\alpha(x))_{\alpha \in \mathcal{A}^k} \tag{2.1}$$

where $\alpha$ denotes all possible subsequences of length $k$ in the sequence $x \in \mathcal{X}$ and $\phi_\alpha(x)$ is the number of times $\alpha$ occurs in $x$.

Using a feature map (Equation 2.1), input sequence $x$ is implicitly transformed into the vector of $\mathbb{R}^{l^k}$. Each coordinate of a vector is indexed by a $k$-mer $\alpha$ and capture the frequency of $\alpha$ in $x$. Therefore, without building a complex model such as multiple sequence alignment or profile HMM, spectrum information of sequences can be extracted as a form of vectors in the *"$k$-mer feature space"*. Then, the $k$-spectrum kernel, the similarity of two sequences $x$ and $y$, is measured in the $k$-mer feature space using the inner product (Leslie *et al.*, 2001).

$$K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle \tag{2.2}$$

The $k$-spectrum kernel measures similarity by co-occurring $k$-mers in the data. The similarity value increases as two sequences $x$ and $y$ contain more common $k$-mers.

Based on the $k$-spectrum kernel, distance is defined as (Leslie *et al.*, 2004):

$$\begin{aligned} \widetilde{K}_k(x, y) &= \frac{K_k(x, y)}{\sqrt{K_k(x, x)}\sqrt{K_k(y, y)}} \\ D_k(x, y) &= \sqrt{\widetilde{K}_k(x, x) + \widetilde{K}_k(y, y) - 2\widetilde{K}_k(x, y)} \end{aligned} \tag{2.3}$$

between two sequences $x$ and $y$. Although this similarity method is less accurate or effective than BLAST (Altschul *et al.*, 1990; Gish and States, 1993) or Smith-Waterman (Smith and BEYER, 1976), it does not require sequence alignments, so it is inexpensive and allows comparison of variable length sequences. For this reason, the $k$-spectrum kernel has been extended in various ways: weighted sum of $k$-spectrum kernel with different $k$ (Smola and Vishwanathan, 2003), considering $m$-mismatches when counting occurrences of $k$-mers (Leslie *et al.*, 2004), combination of count vector and statistical background models (Song *et al.*, 2013; Allman *et al.*, 2017), a weighted sum of hashed $k$-mers by information contents (Murray *et al.*, 2017).

### 2.2.2 The ranked $k$-spectrum string kernel with a landmark

Here, I introduce the ranked $k$-spectrum string (RKSS) kernel. It is an extension of the $k$-spectrum string kernel. Keeping advantages of the string kernel, two features are added for comparative and evolutionary comparison:

- Build and use a common $k$-mers template to encapsulate information of a common ancestry and

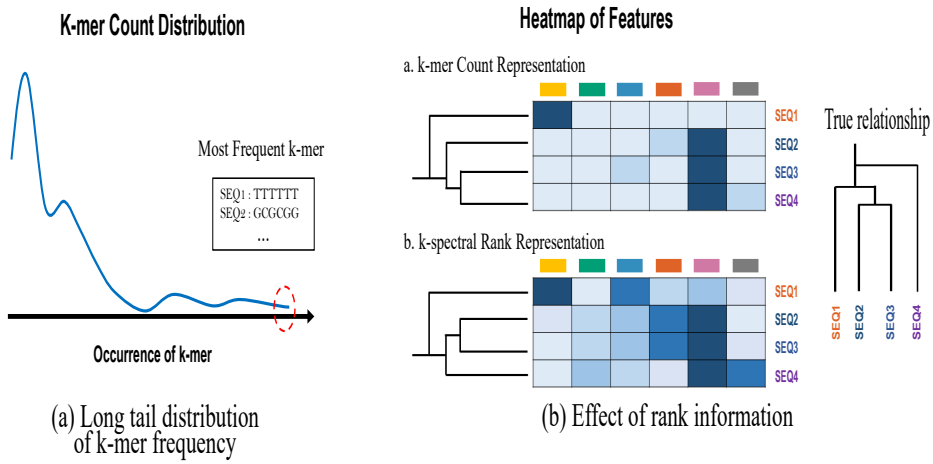



**Figure 2.2:** The efficiency of rank information on genome-scale sequence analysis. (a) Frequency vector of $k$-mer from a sequence encapsulates the information in the original sequence. But when dealing with a large-scale genome sequence, frequency of particular $k$-mer can spike due to repetitive elements or copy number variations. Thus, the distribution of $k$-mer count takes a long-tail distribution form. (b) If frequencies of particular $k$-mers are abnormally high (ex. pink-color/the fifth $k$-mer panel), rank information can more accurately capture subtle distance differences between genomes. For example, consider four sequences (SEQ1 to 4) with a depicted true relationship. In the heatmap representation of frequency of $k$-mers, in the case of using count directly (upper heatmap), the pink-color $k$-mer panel (the fifth $k$-mer)'s frequency is unexpectedly high and does not reflect the subtle distance difference between SEQ2, 3, and 4. On the other hand, when rank information is used (lower heatmap), the bias problem of count is removed, and the relationship between the SEQs is more accurately detected.

- Use of correlation in ranks of $k$-mers instead of occurrence counts

Differences of the conventional $k$-spectrum string kernel and the proposed RKSS kernel are illustrated in Figure 2.2(b) and (c).

The RKSS kernel is defined as follows:

$$K_k^{Rank}(x, y) = RC(\Phi_k^{common}(x), \Phi_k^{common}(y)) \tag{2.4}$$

where $\Phi_k^{common}$ is a feature map on the common template or landmark of $k$-mers and $RC$ is the Kendall tau rank correlation. It measures a concordance of each rank pair between two sequences. As the number of concordant rank pairs increases, the value of the ranked kernel increases.

Using the RKSS kernel (Equation 2.4), a kernel distance is defined within two sequences $x$ and $y$ as:

$$\begin{aligned} \widetilde{K}_k^{Rank}(x, y) &= \frac{1 + K_k^{Rank}(x, y)}{2} \\ dist(x, y) &= \sqrt{\widetilde{K}_k^{Rank}(x, x) + \widetilde{K}_k^{Rank}(y, y) - 2\widetilde{K}_k^{Rank}(x, y))} \\ &= \sqrt{1 - K_k^{Rank}(x, y)} \end{aligned} \tag{2.5}$$

where $\widetilde{K}_k^{Rank}(x, x) = 1$ with self-similarity property. Using the kernel distance, a pairwise distance between sequences or genomes is calculated.

The reason for using a common template or landmark is to capture differences among genome sequences. As mentioned earlier, the string kernel is a technique for extracting information contained in a biological sequence by an alignment-free manner. This is obviously inexpensive than an alignment-based approach. However, this technique has difficulty in capturing functional, structural and/or evolutionary relationships between sequences while alignments methods can easily handle these relationships. To overcome this problem, a landmark is utilized for the kernel.

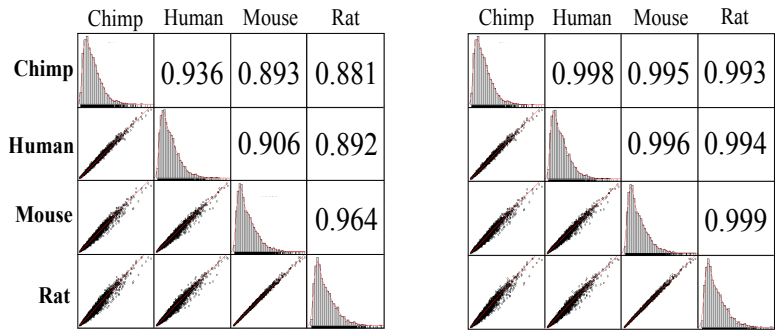
*Comparison of the RKSS kernel and the spectrum kernel:* To further support the power of RKSS kernel for comparative and evolutionary comparison,

an additional experiment were performed to compare the distance between genomes using the RKSS kernel and the spectrum kernel. Four genomes such as human, chimp, mouse, and rat were compared. The goal of the test was to see how well four genomes were separated when distances among the four genomes were computed using the RKSS kernel and the spectrum kernel. To compute distance using the RKSS kernel, a similarity between two genomes was computed by Kendall rank correlation and it was converted to a distance by the kernel method (Equation 2.5). Likewise, to compute distance using the spectrum kernel, a similarity between two genomes was computed by the inner product and it was converted to a distance by the kernel method (Equation 2.3). The results of pairwise genome similarity using the RKSS kernel and the spectrum kernel were summarized in Figure 2.3. As shown in the figure, similarities between two distant groups become bigger when the Kendall rank correlation was used. This means that difference in pairwise genome distances measured by the RKSS kernel gets bigger than the spectrum kernel. This experiment supports that the RKSS kernel is more effective than the spectrum kernel for comparative and evolutionary genome comparison.

In order to demonstrate efficiency of the RKSS kernel in comparative and evolutionary studies, two application studies were designed by using the RKSS kernel with the single(multiple) landmark(s) concept.

### 2.2.3 Single landmark-based reconstruction of phylogenetic tree

The first application is the reconstruction of a phylogenetic tree with a single landmark (Figure 2.1(b)). The main question in this experiment is to model the evolution times between species on genome sequence level, which contains repetitive elements or copy number variations. This problem may be addressed by the single landmark that it represents a hidden common ancestor of all species and pair-wise similarities between species are determined by the land-




(a) Similarity by RKSS kernel (b) Similarity by Spectrum kernel

**Figure 2.3:** Similarity comparison of (a) the RKSS kernel and (b) the spectrum kernel. Four species with two groups (Chimp and Human vs Mouse and Rat) were analyzed for comparison of two kernel methods. The 6-mer distributions were represented at diagonal part of the figure. Target genome region was exon. Based on the similarity between distant species, the RKSS kernel captured the relatively subtle distance across species better than the spectrum kernel.

mark. To elucidate relationships between 10 mammalian species, three types of sequences are utilized such as exon, intron, and CpG island sequences.

In addition to the RKSS kernel, the conventional $k$-spectrum string kernel is used for comparison in this phylogenetic tree reconstruction experiment. In case of RKSS kernel, a common $k$-mer template of 10 species, single landmark, is built by frequencies of all $k$-mers among all species. Then, pair-wise similarities and distances of species are calculated by Equation 2.4 and 2.5. On the other hand, the $k$-spectrum string kernel measures similarities between species pair-wisely by using all $k$-mers frequency vectors and Equation 2.2. The distance matrix is generated using those similarities by Equation 2.3.

Then, a neighbor joining (NJ) algorithm (Saitou and Nei, 1987) is used to reconstruct phylogenetic trees using distance matrices from two string kernel methods. NJ is a method of a distance-based tree structure that can eliminate

errors that may occur with Unweighted Pair Group Method using Arithmetic Average (UPGMA) method. While UPGMA looks for nearby nodes based on distance, NJ tries to find a neighbor set that minimizes overall tree length along with it. Especially, it is known that NJ makes a reasonable tree nearer the evolutionary distance.

### 2.2.4 Multiple landmark-based distance comparison of exons, introns, CpG islands

For the second application to show the usefulness of RKSS kernel, a comparative analysis is performed among genomic regions such as exons, introns, and CpG islands. The main question in this experiment is to compare exons, introns, and CpG islands in terms of distances, which can show reveal similarities between these three regions in a biological context. This investigation is possible since all three types of sequences are mapped into a single feature space. For this comparative study of 10 mammalian genomes, three landmarks for exons, introns, and CpG islands for each genome are constructed.

This approach is inspired by the work in (Middleton and Kim, 2014). In the study, authors performed clustering of RNA structures without folding or alignment. The core of this approach is to calculate relative distances between two sequences on a feature space. RNA family covariance models (Rfam CMs) were obtained from the Rfam database (Burge *et al.*, 2012) and the distances between input sequences and Rfam CMs were calculated. Then, a new feature space was constructed by assigning each dimension as Rfam CM. More specifically, a sequence $x$ has changed to a vector in a new feature space, in which the coordinate index by Rfam CM will be the distance between Rfam CM and $x$. Rfam CMs used in this process are defined as "landmark" and RNA structure clustering is performed using distances between input sequences on the newly defined landmark space.

Figure 2.1(c) shows the illustration of how to construct the landmark space for three genomic regions and details of the method are explained step by step below.

1. Input: the species set $S$ and all three genomic regions (exon, intron, and CpG island)

2. Find the common $k$-mer template among the species by RKSS kernel in each of the three regions.

3. Using the common template of each region, obtain a rank vector of each region of each species and name it landmark of the region in the species. The landmark space $L \subset \mathbb{R}^{3|S|}$ is constructed by those landmarks (The number of dimension = The number of landmarks = The number of genomic regions $\times$ The number of species).

4. Given a sequence $x$, measure a similarity with the landmarks by RKSS kernel (Equation 2.4) and calculate a kernel distance by Equation 2.5.

5. The sequence $x$ is transformed into a vector of landmark space as follows: for each landmark $l$, the coordinate indexed by $l$ will be a kernel distance between a landmark $l$ and a sequence $x$ as described in step 4.

The following analyses are performed on the constructed landmark space. 1) Whether three regions in the landmark space are distinguished. 2) What is the correlation between landmarks 3) Whether it is possible to assign a correct region to a given unknown sequence.

### 2.2.5 Sequence Data for analysis

From the UCSC Genome Browser database (Kent *et al.*, 2002), sequences of three regions (exon, intron, and CpG island) in 10 mammalians were downloaded using table browser program in the UCSC and utilized for the analysis.

**Table 2.1:** The List of 10 mammalian species. In the UCSC Genome Browser, the sequence data of three genomic regions were downloaded from the genome assembly corresponding to each data version. The number of sequences varies depending on the species and region

| Species | Data version | exon | intron | CpG island |
|---|---|---|---|---|
| Chimp | CSAC 2.1.4/panTro4 | 2,105 | 1,988 | 28,310 |
| Cow | Bos_taurus_UMD_3.1.1/bosTau8 | 13,638 | 13,221 | 37,226 |
| Dog | Broad CanFam3.1/canFam3 | 1,718 | 1,578 | 48,192 |
| Human | GRCh38/hg38 | 56,198 | 68,294 | 30,477 |
| Marmoset | WuGSC 3.2/calJac3 | 219 | 217 | 32,732 |
| Mouse | GRCm38/mm10 | 32,889 | 34,180 | 16,023 |
| Opossum | Broad/monDom5 | 351 | 227 | 22,441 |
| Pig | SGSC Sscrofa10.2/susScr3 | 4,921 | 4,464 | 43,643 |
| Rat | RGSC 6.0/rn6 | 18,218 | 16,384 | 18,218 |
| Rheus | BCM Mmul_8.0.1/rheMac8 | 5,832 | 5,418 | 30,560 |

Table 2.1 shows 10 mammalian reference genomes and their versions from sequences are taken.

## 2.3 Results

### 2.3.1 Reconstruction of phylogenetic tree on the exons, introns, and CpG islands

The goal of this experiment was how well each of exons, introns, and CpG islands could construct phylogenetic trees when RKSS kernels were used. Sequences from the three regions on the genomes were collected as described in Table 2.1 and reconstructed phylogenetic trees respectively. Three regions selected for analysis are very widely distributed on the genome, and the sequence

**Table 2.2:** List of the mitochondria gene of 10 mammalian species. Those genes are used for reconstruction of reference phylogenetic tree for comparison of the RKSS kernel and the $k$-spectrum string kernel. Data can be downloaded in the URL as form of "https://www.ncbi.nlm.gov/nuccore/<GenBank Accession>/<GeneBank Accession>"

| Species | GenBank Accession |
|---|---|
| Chimp | X93335.1 |
| Cow | AF492351.1 |
| Dog | AY729880.1 |
| Human | V00662.1 |
| Marmoset | NC_021941.1 |
| Mouse | NC_005089.1 |
| Opossum | Z29573.1 |
| Pig | AF486855.1 |
| Rat | X14848.1 |
| Rheus | AY612638.1 |

lengths and numbers are very different. For these reasons, it was difficult to distinguish species and reconstruct phylogenetic trees using alignment-based approaches. Since methods that could be used in this situation are alignment-free methodologies, $k$-spectrum string kernel, which was the most similar to RKSS kernel and was a basis of the string kernel method, was used for comparison of reconstructing power of phylogenetic trees.

For a better comparison of two kernels, an additional ground truth-like phylogenetic tree was built from mitochondrial genomes that were more conserved and refined than above three regions (van de Sande, 2012; Li *et al.*, 2013; Zubaer *et al.*, 2018). To construct a phylogenetic tree of 10 mammalians, the

mitochondria genomes of all species were collected (Table 2.2). CLUSTALW (Larkin *et al.*, 2007), a multiple sequence alignment tool, was used to further clarify the inter-species comparative analysis from the collected sequences. After pair-wise distances were calculated, the tree was reconstructed by the neighbor joining algorithm.

Figure 2.4 showed the phylogenetic trees of three regions reconstructed by two kernel methods as well as the mitochondrial genome tree (MT tree). All trees were drawn by online phylogenetic tree visualization tool (PhyIO (Robinson *et al.*, 2016)). Both of kernel methods used $k$-mers of length 6. More specifically, Figure 2.4(b) to (d) were phylogenetic trees reconstructed by RKSS kernel method using top 100 common 6-mers of each region as one landmark; exon, CpG island, and intron respectively. As the same order, Figure 2.4(e) to (g) were phylogenetic trees reconstructed by the $k$-spectrum kernel method. Unlike the RKSS kernel method, this method used all possible 6-mers, i.e., 4,096 $k$-mers, to measure pair-wise distances between species.

As a reference tree for comparison, three notable groups were spotted in the phylogenetic tree from mitochondrial sequences. The first group (names as MT1; red group) contained Human, Chimp and Rhesus. The second group (names as MT2; yellow group) contained Mouse and Rat. The last group (names as MT3; green group) contained Pig, Cow, and Dog. These formations have also been reported in previous studies (Miller *et al.*, 2007; Huising *et al.*, 2006; Sequencing *et al.*, 2014). When comparing the reconstruction results of the two kernel methods, these three groups were used as the main criteria.

Let start with comparisons of two kernel methods on the exon region (Figure 2.4(b) and Figure 2.4(e)). Results showed that the two kernel trees are similar to the MT tree. In both kernel methods, all three MT groups are well clustered. However, subtle differences could be found. First, in the case of the spectrum kernel, cluster formation of Pig, Cow, and Dog was consistent

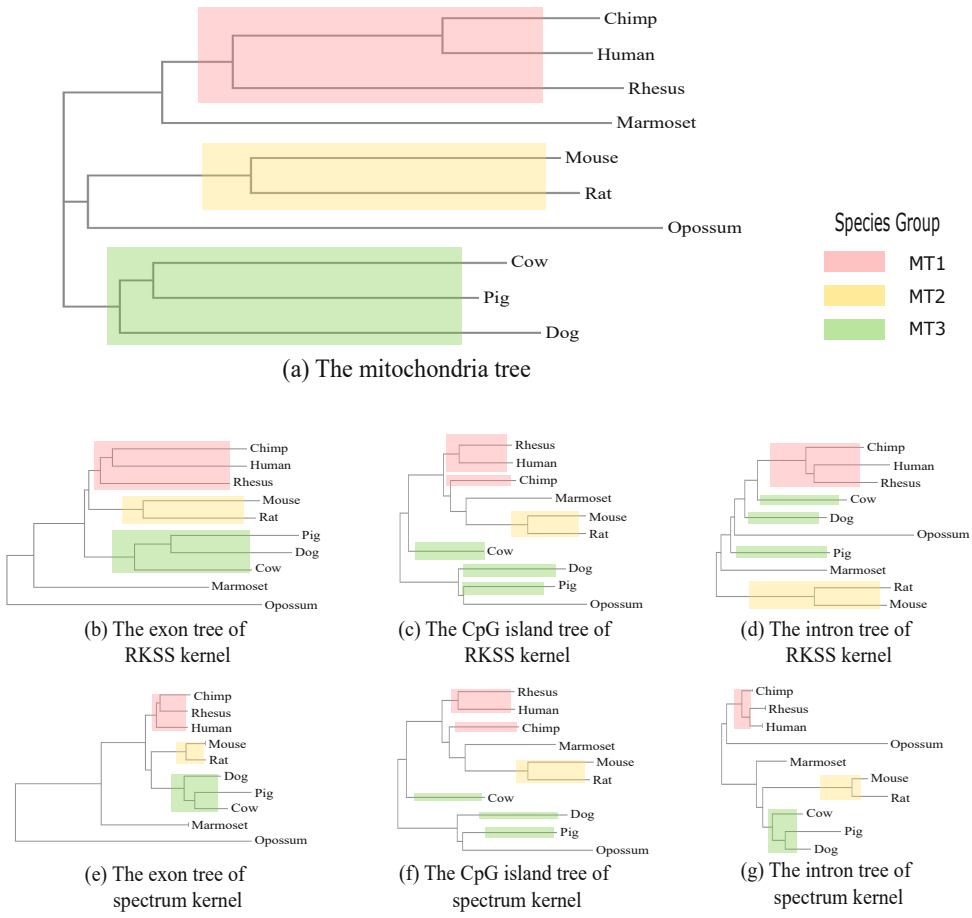



**Figure 2.4:** Comparison of phylogenetic tress of two kernel methods. (a) Phylogenetic tree of the mitochondria gene tree. It is generated from result of multiple sequence alignment by CLUSTALW. Overall structure of tree is consistent with previous studies. (b)-(d) Phylogenetic trees of RKSS kernel method on exon, CpG island, and intron, respectively. Length of $k$-mer is 6 and top 100 common 6-mers across 10 species are selected as one landmark on each region. (e)-(g) Phylogenetic trees of the $k$-spectrum string kernel method on exon, CpG island, and intron, respectively. Length of the $k$-mer is 6.

with the MT3 group. However, the relationship between Human, Chimp, and Rhesus was not properly reflected with MT1 group. When comparing DNA sequences, Human and Chimp showed 98% to 99% similarity, while Human and Rhesus showed 93% similarity. Based on these points, it was reasonable to assume that Human and Chimp were the most similar, and followed by Rhesus, as shown in the MT tree. This relationship was correctly captured by RKSS kernel, but not by the spectrum kernel. One more notable difference was the clustering order of MT groups. In case of RKSS kernel, as shown in Figure 2.4(a) and (b), all MT groups were correctly clustered with proper order. It was also consistent with previous studies (Miller *et al.*, 2007; Huising *et al.*, 2006; Sequencing *et al.*, 2014). However, the $k$-spectrum kernel failed to reconstruct this formation and shows the cluster of MT2 group and MT3 group. Considering these points, the exon tree created by RKSS kernel was well created and better than those of the $k$-spectrum kernel.

Similar comparisons were performed on the CpG island and intron region. Two kernel methods reconstructed the same tree of CpG island (Figure 2.4(c) and (f)). This result was interesting in a sense that RKSS kernel only utilized 100 $k$-mers with a single landmark, whereas $k$-spectrum kernel utilized all possible $k$-mers. This indicated that even one landmark information used as a reference point between the species was enough in calculating pair-wise distances of species. Although the tree made with CpG island had poor performance compared to the tree constructed from the exon region, the overall form was still quite reasonable in the biological sense. This suggested that the CpG island also contains evolutionary information that could be utilized in distinguishing the species (Chae *et al.*, 2013).

In the case of intron trees (Figure 2.4(d) and (g)), two trees looked totally different and also were dissimilar with the general tree patterns of exon and CpG island. In the case of RKSS kernel tree (Figure 2.4(d)), species belonging

to MT1 and MT2 group were well clustered. But species of MT3 group were broken and clustered with other species. In addition, Opossum interfered with MT1 and MT2 groups. On the other hand, the intron tree generated by the spectrum kernel looked pretty good except for the binding position of Opossum and composition of MT groups. On close inspection, however, the tree did not make sense with respect to evolutionary time modeling. While other five trees preserved the divergence time of each species relatively well corresponding to common knowledge on phylogeny, this tree failed in reproducing such evolutionary time as primate group was determined to be closest to a hidden common ancestor. This pattern was obviously different from results from other trees and was in conflict with common knowledge on evolution. Thus, it was premature for us to conclusively compare intron trees from two methods and to tell which result was better.

As shown in Figure 2.4, RKSS kernel succeed in capturing evolutionary information relatively well compared to the widely used $k$-spectrum kernel. This pattern was also observed in additional experiments of phylogenetic tree reconstructions that were performed with different $k$-mers and top common $k$-mers ($k$ = [3, 4, 5] and topN = [64, 100, 200, 500, 1000]). Furthermore, the RKSS kernel reconstructed more reliable trees than other distance methods such as Euclidean distance and Jensen-Shannon divergence (Figure 2.5). From those experiments and frequency distribution of common $k$-mers, a long-tail like distribution, 3-mer or 6-mer were recommended for the RKSS kernel, which can reflect biological knowledge such as codon and dicodon. Also, in the case of the number of common $k$-mers, it would be better to look at the frequency distribution of the common $k$-mer to select the number that can reflect the characteristics of the data with respect to the number of features.

On the other hand, during the performance comparison of the two kernels in three regions, it was found that interesting property about exons, introns,

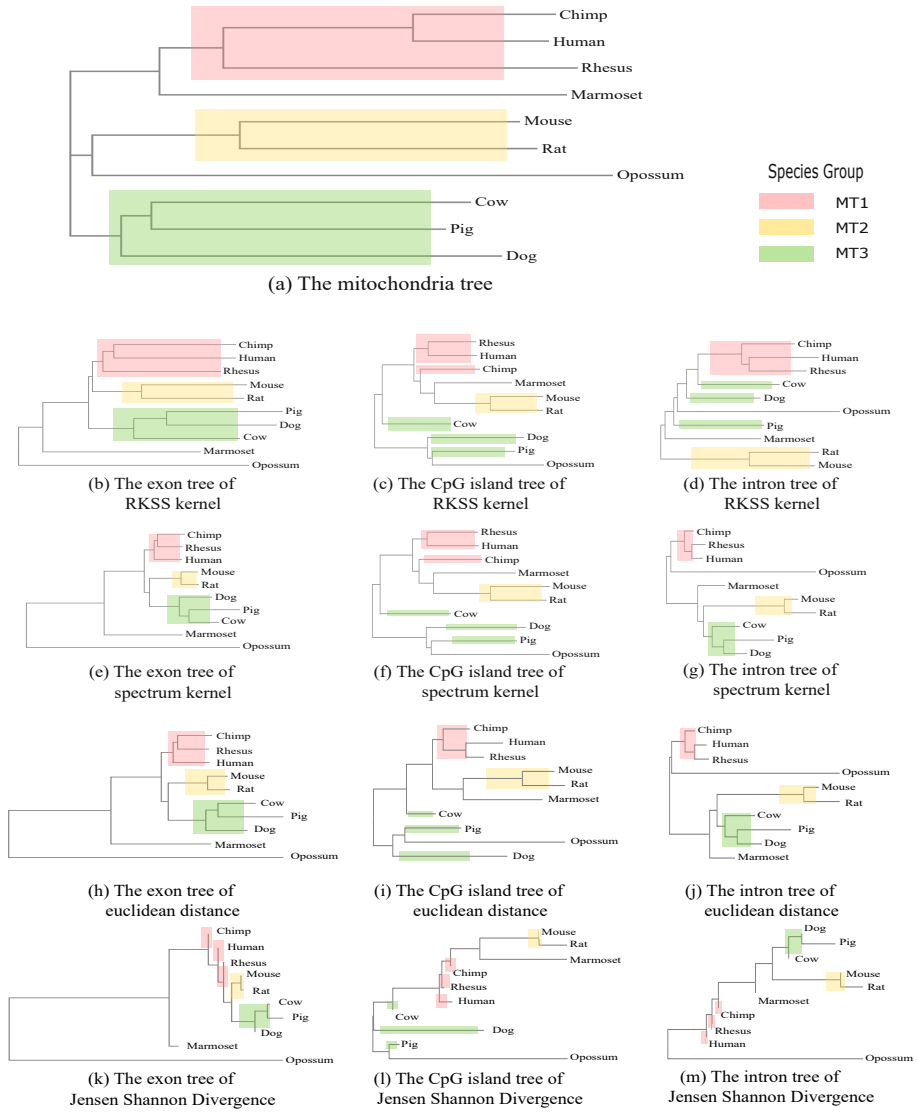



**Figure 2.5:** Phylogenetic tree comparison on 6-mer with Euclidean distance and Jensen-Shannon divergence. ((a) Phylogenetic tree of the mitochondria gene tree. It is generated from result of multiple sequence alignment by CLUSTALW. Overall structure of tree is consistent with previous studies. (b)-(d) Phylogenetic trees of RKSS kernel method on exon, CpG island, and intron, respectively. (e)-(g) Phylogenetic trees of the $k$-spectrum string kernel method. (h)-(j) Phylogenetic trees of the Euclidean distance method. (k)-(m) Phylogenetic trees of the Jensen-Shannon divergence method.

and CpG islands in terms of evolutionary information content. Investigation of the trees generated by RKSS kernels, the exon tree was most similar to the MT tree, followed by the CpG island tree and the intron tree. This indicated that there was a lot of evolutionary information in the exon region, whereas the intron region had a relatively small amount of information. Considering that the reconstruction performance of the CpG island tree was between those of the exon tree and the intron tree, it could be expected that the amount of evolutionary information in the three regions was in the order of exon > CpG island > intron.

### 2.3.2 Landmark space captures the characteristics of three genomic regions

In the previous section, the RKSS kernel of exons had enough information to reconstruct the phylogenetic tree of MT sequences well while the RKSS kernel of introns did have relatively small information to reconstruct the phylogenetic tree. CpG islands stood in between exons and introns in terms of the ranked $k$-spectrum feature space. In this experiment, exons, introns, and CpG islands were put into different feature spaces with respect to one landmark $k$-mers. Thus, in this section, I performed an experiment where all exons, introns, and CpG islands were put into a single space rather than separate feature spaces. The goal of this experiment was to compare exons, introns, and CpG islands in terms of kernel distances.

To achieve this goal, multiple landmarks were used and a new feature space (named as landmark space) was constructed through the process mentioned in Section 2.2.4. A total of 30 landmarks were generated from 10 mammalian species and three genomic regions (exon, intron, and CpG island). If one landmark aimed to capture evolutionary information of the region and identify pair-wise distances between species, multiple landmarks were used to cap-

ture the characteristics of exons, introns, and CpG islands commonly found in species. More specifically, it elucidated hypothesis about genomic or evolutionary information content in the exon, intron, and CpG island that was found in phylogenetic tree reconstruction experiment: exon > CpG island > intron.

Based on the hypothesis, if a sequence contained a lot of genetic information, the sequence was close to the landmarks of exon family and might be located farthest away from the intron landmarks. Observations of this objective could be applied similarly to other regions too. To demonstrate the hypothesis on the landmark space, a information theoretic concordance test on rank was performed (Figure 2.6). For that, a template rank vector was made; for example, coordinates of the exon-based landmarks had high ranks, those of the CpG island-based landmarks had intermediate ranks, and those of the intron-based landmarks had low ranks. In a similar manner, when a sequence was mapped to the landmark space, each coordinate of a feature vector of the sequence was indexed by RKSS kernel distance between the sequence and each landmark. In detail, the kernel distance between sequence and landmark was calculated using the feature map and the rank profile of landmark by Equation 2.4 and 2.5.

Figure 2.7 showed the concordance test results of 6-mer landmarks with 10 mammalian species. Sequences of each region were individually mapped to the landmark space and concordance tests were performed. Average values of concordance with the hypothesis by region were represented as a bar plot. As shown in Figure 2.7, except for Opossum, average values of concordance with the hypothesis in the other nine species showed this ordering: exon > CpG island > intron.

As expected, exon sequences showed the highest concordance, indicating that genomic information of exon sequences was higher than those in the other

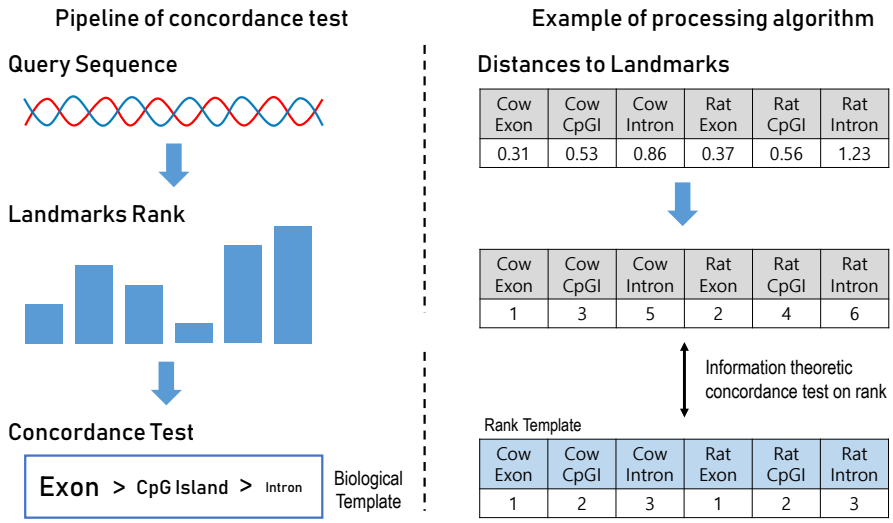



**Figure 2.6:** Example of how to measure the concordance value of each sequence on the landmark space. Given a query sequence that does not know which genomic region it belongs to, it was mapped on the landmark space. Each coordinate was indexed by the RKSS kernel distance (Equation 2.5) between a query sequence and the corresponding landmark. These distances were converted into the rank which smaller distances had higher ranks (ascending order). Finally, based on the rank concept, the information theoretic concordance was calculated between a query sequence rank vector and the rank template of hypothesis to be verified.

two regions. The amount of information contained in the other two sequences was also considered to agree with the hypothesis. This observation was also demonstrated in experiments with different $k$s ($k$=[3, 4, 5]).

The results showed the hypothesis about genomic information contents of exon, intron, and CpG island. Understanding what makes the differences in information contents between the three regions was important. Also, it was worth to figure out the reason for the weird pattern of exon sequences in Opossum. In order to address these questions, correlations between landmarks

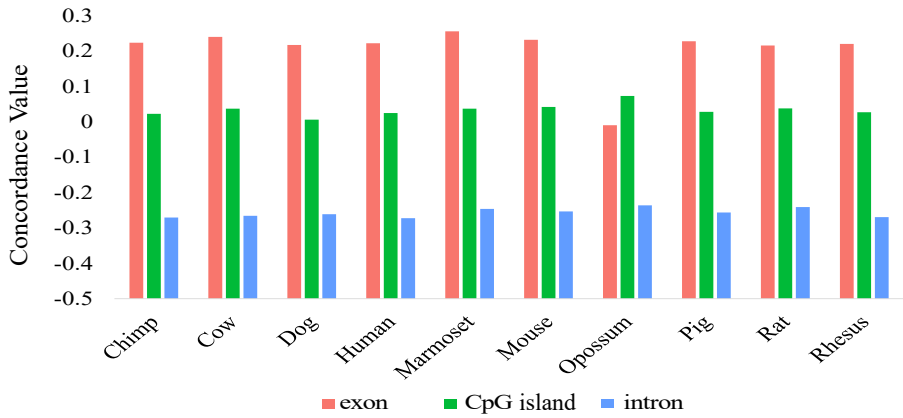
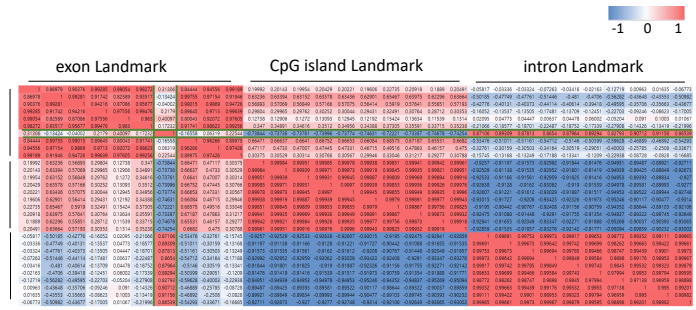



**Figure 2.7:** Concordance test for the three region and 10 species with the 6-mer landmark space. For each sequence mapped to the landmark space, a concordance test with the following hypothesis was established: Order of Information contents belonging to the region was exon > CpG island > intron. Concordance was tested with the information theoretic concordance test between the template rank vector of the hypothesis and the landmark space vector of the sequence. The average concordance value of each region was expressed in a bar graph. In the nine species except for Opossum, the concordance order was consistent with the hypothesis.

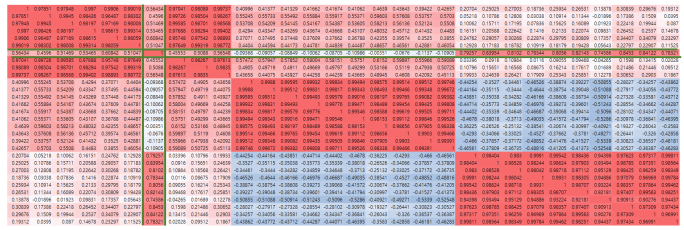
were analyzed on two $k$-mers, 3-mer and 6-mer, because the two $k$-mer formed biologically important codon and dicodon, respectively. Figure 2.8 and 2.9 showed the heatmap of the correlation between the values of each landmark dimension when the exon, CpG island, and intron sequences of Chimp were mapped to landmark space. In Figure 2.8, the same features were observed, no matter where the sequence of any region was mapped. The exon landmark and CpG island landmarks showed a positive correlation. The intron landmarks, on the other hand, showed a negative correlation with the other two regions, especially strongly negative correlation with CpG island landmarks. These

features were consistent with the fact that the value of the exon region and the value of the CpG island region in the concordance test performed in the landmark consisting of 3-mer did not differ greatly.

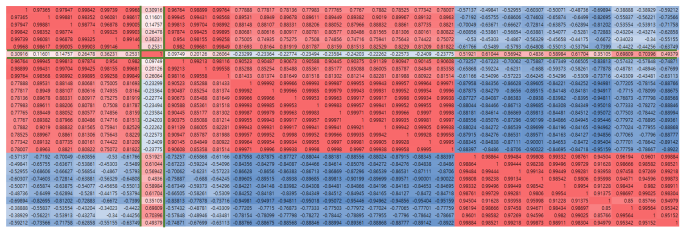
In the case of k=6, the patterns of correlation relations were slightly changed in Figure 2.9. In both cases of mapping the exon sequence to the CpG island sequence, the correlation between landmarks between different regions had weakened against to result of k=3. In addition, patterns of heatmap were very similar when mapping the exon sequences and the CpG island sequences. From these properties, it was possible to interpret the reason why the differences of concordance values were in-creased in all species (by the reduced correlation between exon landmarks and CpG island landmarks). This was because dicodon contains larger amounts of information than the codon, RKSS kernel could capture the difference in the amount of information hidden in the different $k$-mers.




(a) Mapping exon Sequences

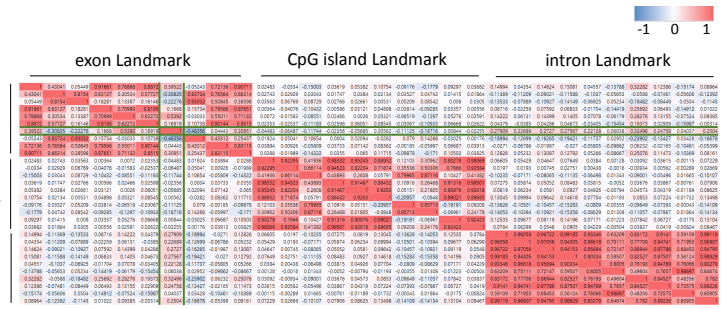


(b) Mapping CpG island Sequences



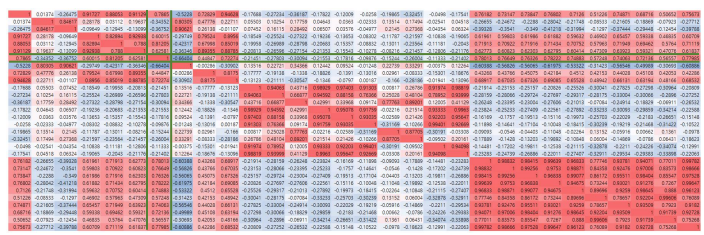
(c) Mapping intron Sequences

**Figure 2.8:** The heatmap of correlation between landmarks when the Chimp sequences were mapped into the 3-mer landmark space. The heatmap was symmetric and the order of columns(rows) was exon, CpG island, intron landmarks and each species were sorted as the alphabet order. (a) The result of correlation between landmarks when Chimp exon sequences were mapped. Landmarks in each region showed a strong positive correlation only with each other, and there was little positive/negative correlation with CpG island landmarks/intron landmarks. (b) The result of correlation between landmarks when Chimp CpG island sequences were mapped. The pattern similar to (a) was observed. (c) The result of correlation between landmarks when Chimp intron sequences were mapped. Weak negative correlations between intron landmarks and exon landmarks, as well as strong negative correlations with CpG island landmarks, were observed.




(a) Mapping exon Sequences

(b) Mapping CpG island Sequences



(c) Mapping intron Sequences

**Figure 2.9:** The heatmap of correlation between landmarks when the Chimp sequences were mapped into the 6-mer landmark space. (a) The result of correlation between landmarks when Chimp exon sequences were mapped. Landmarks in each region showed a strong positive correlation only with each other, and there was little or no correlation between landmarks in other regions. (b) The result of correlation between landmarks when Chimp CpG island sequences were mapped. The pattern similar to (a) was observed. (c) The result of correlation between landmarks when Chimp intron sequences were mapped. Weak positive correlations between intron landmarks and exon landmarks, as well as weak negative correlations with CpG island landmarks, were observed.

On the other hand, there were some anomalies common in Figure 2.8 and 2.9. It showed the opposite correlation with other landmarks, the cause of which was Opossum exon landmark (high-righted in the figures by green lines). Looking at the correlation values, the Opossum exon landmark seemed to show a weak correlation with other exons. However, there was a stronger positive correlation with intron landmarks. It also had a negative correlation with CpG island landmarks. Based on this, the exon sequences of Opossum might have $k$-mer distributions close to the intron sequences. This indicated that why the value of the exon in the concordance test was low.

### 2.3.3 Cross-evaluation of the landmark-based feature space

To confirm robustness of the landmark space, a further experiment was performed if the landmark space could produce the same results in Section 2.3.2 when sequences in the unknown region were given as inputs. If the landmark space correctly reflected characteristics of exons, introns, and CpG islands, regional differences will be identified even for unknown sequences. Experiments were conducted in a cross-validation-like manner where one of the species was selected as test data. Other species were used for constructing the landmark space.

The experiment showed the concordant test of cross-evaluation of the landmark space in the same manner. As a result of concordance test, comparison of information contents in exons, introns, and CpG islands matched well with the previous result: exon $>$ CpG island $>$ intron. Frankly, compared to Figure 2.7, the concordance values were slightly decreased. However, the patterns of values were consistent with the hypothesis. It implied that a landmark space properly captured the characteristics of exons, intron, and CpG islands.

# Chapter 3

# Pathway-based cancer subtype classification and interpretation by attention mechanism and network propagation

Genes perform biological functions through interactions with other genes. An effective way to analyze complex genetic interactions is to use biological networks. The most widely used knowledge of biological networks is biological pathways such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). KEGG consists of hundred small networks, each of which is designed to represent distinct biological process. In this study, I study the interaction between genes by using pathway information and develop a model for predicting cancer subtype. By extracting information from each pathway and generating an interpretable level of results in which pathways are useful, biological phenomena specific to cancer subtypes can be analyzed.

## 3.1 Motivation

Biological systems are too complex to understand as a whole. For this reason, biological systems are dissected to small subsystems that can be easily understood. The most widely used subsystems are biological pathway databases that are curated for years and these pathway databases such as KEGG (Kanehisa and Goto, 2000) are widely used to analyze transcriptome data.

Cancer subtypes are often classified based on gene expression profiles. For example, breast cancer is well characterized in terms of subtypes that are widely used for clinical applications (Grimm *et al.*, 2014; Hwang *et al.*, 2019). However, cancer subtype classification based on gene expression profiles showed poor stability on independent datasets (Kim *et al.*, 2012; Alcaraz *et al.*, 2017). Furthermore, it does not provide insightful biological information such as subtype-specific activations of certain pathways (Gatza *et al.*, 2010; Segura-Lepe *et al.*, 2019). Thus, pathway-based cancer subtype classification is desirable since pathways can be an effective way to generate landscape of molecular functions of an organism as a collection of biological knowledge (Viswanathan *et al.*, 2008). While pathway databases contain static information in general, mapping transcriptome data to the pathways can enhance their usefulness by explaining the dynamics of cancer in terms of biological functions (Schadt *et al.*, 2005; Kunz *et al.*, 2019). Another view on why pathway-based cancer subtype classification is useful can be explained in terms of the number of dimensions or variables since the dimensionality from genes to pathways is two orders of magnitude smaller (20000 vs. 300), resulting in better interpretability on the feature space (Glaab *et al.*, 2010; Gatza *et al.*, 2010; Su *et al.*, 2009). However, there is a serious issue when pathways are used for cancer subtype classification. Only a fraction, 1/3 in the case of human, is included in biological pathways and use of pathways is limited in predictive

power for subtype classification, compared to use of the entire transcriptome. Thus, the main research question here is:

**How can pathways be used effectively for cancer subtype classification?**

Most of pathway analysis tools are developed to measure pathway activation levels (Lim *et al.*, 2018). Given that each pathway is modeled as a single value of representing the activation status of the pathway, combining these values for the entire biological system is not straightforward, often resulting in poor performances of cancer subtype classification (Lim *et al.*, 2018).

In this study, a deep learning approach is proposed to investigate three important research questions.

1. How can accuracies be improved in predicting cancer subtypes using transcriptome data in terms of pathways?

2. How different are pathway interactions among cancer subtypes?

3. Why are gene expression profiles different among cancer subtypes?

To begin with modeling individual pathway, an effective computational method that can consider interactions among genes is needed. Due to recent advances in deep learning, *graph convolutional network (GCN)* can handle these interactions instead of traditional pathway analysis tools (Defferrard *et al.*, 2016; Kipf and Welling, 2017). Aggregation of node features (= gene expression levels in this study) on the graph are performed in various ways such as spectral graph convolution (Dhillon *et al.*, 2007; Defferrard *et al.*, 2016), layer-wise propagation (Kipf and Welling, 2017), diffusion process (Atwood and Towsley, 2016; Monti *et al.*, 2017), graph embedding for sparse connections between nodes (Kong and Yu, 2018). Like convolutional neural network (CNN), GCN can capture localized patterns in data, and unlike CNN, it can

be used for non-grid structured data such as graph. For these reasons, GCN has been successfully used in protein-protein interaction (PPI) network for the prediction of breast cancer subtype and drug side effect (Rhee *et al.*, 2018; Zitnik *et al.*, 2018).

Given a GCN model for each pathway, interactions between pathways can be considered as a network of pathways by combining several hundred pathway models again. For example, a condition-specific pathway network can be built from transcriptome data (Moon *et al.*, 2017) and GCN can be used again for combining several hundred pathway models. However, GCN is a deep learning model which is black-box model that cannot explain which input features are important and why the model performs well (Castelvecchi, 2016). To open up the black-box model, *attention mechanism* is frequently used (Vaswani *et al.*, 2017). The attention mechanism helps identify features that make the models achieve better performances (Choi *et al.*, 2016; Zheng *et al.*, 2017).

Another important question is how to explain the differences between gene expressions and interactions among subtypes in terms of pathways. The question then is how different *biological functions* among cancer subtypes by extending pathway-level information to gene-level (Jo *et al.* (2016)). In this regard, *network propagation* is also widely used in the network analysis for biological interpretation (Pearson, 1905; Cowen *et al.*, 2017). For example, network propagation has been successful in aggregating mutation profiles on the molecular interaction networks to detect significant gene modules (Leiserson *et al.*, 2015; Hofree *et al.*, 2013; Zhang *et al.*, 2018).

In this study, an *explainable deep learning* model (Gunning, 2017) is proposed for cancer subtype classification and pathway modeling. The model consists of three steps (Figure 3.1). To begin with, a pathway model is generated for each of the pathways by GCN to utilize biological prior knowledge. Then, multiple GCN pathway models are integrated into a single model by *multi-*

*attention based ensemble (MAE)*. The MAE model consists of two-level attentions to capture complex pathway combinations of cancer data. Finally, to show how different subtypes are in terms of biological functions, I propose a network propagation method with permutation-based normalization for identification of TFs that influence gene expressions and pathways. In the following sections, Section 3.2 explains detailed implementation of the model. In Section 3.3, the power of the model is demonstrated in experiments with five cancer data sets.

**[Problem Definition of this study]**

Given a set of pathways $i = 1, 2, ..., m$ and gene expression data $X$ with $N$ patients

$< Input >$

$G_i$ : a graph of pathway, $G_i = (V_i, E_i)$

$V_i$ & $E_i$ : a set of genes and interactions in the pathway $G_i$

$X_i$ : a gene expression matrix, $X \in \mathbb{R}^{N \times |V_i|}$

$< Output >$

$Y$ : Cancer subtype of given N patients, $Y = \{0, 1, 2, ..., c\}^N$,

$c$ : number of classes

$< Model >$

Extract pathway information on each pathway $i$

using graph convolutional network (GCN)

Combine the results of GCN by multi-attention based ensemble (MAE)

Predict cancer subtype $Y'$ using $MAE$ model

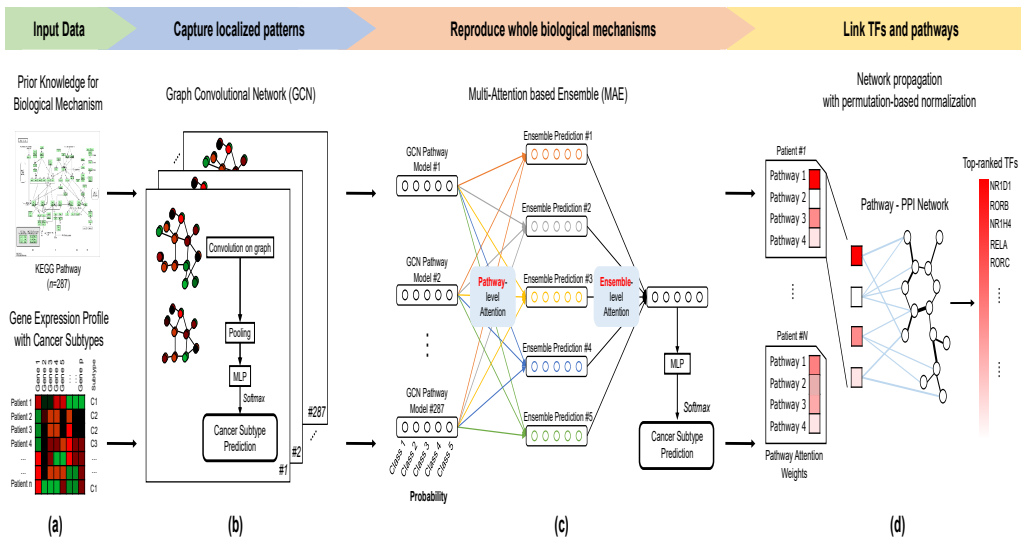



**Figure 3.1:** The workflow of the proposed pathway-based cancer subtype classification model. Given (a) gene expression data and pathway set, the proposed model consists of three major parts. (b) Using graph convolutional networks (GCNs), each GCN pathway model captures localized gene expression patterns. Then, multilayer perceptron (MLP) is followed by GCN to encode extracted gene-level information into pathway level. (c) Multi-attention based ensemble (MAE) combines the outputs of GCN pathway models. To consider heterogeneity of cancer, two attention layers named pathway-level and ensemble-level are utilized. (d) To identify transcription factors (TFs) related to highlighted pathways from (c), network propagation on a pathway-PPI network is performed. To avoid that propagation is over-fitted on the high degree nodes, permutation-based normalization of TFs is considered.

## 3.2 Methods

### 3.2.1 Encoding biological prior knowledge using Graph Convolutional Network

Given a pathway $p$ as prior knowledge, a graph $G^p = (V^p, E^p)$ is determined, where $V^p$ is a set of nodes representing genes and $E^p$ is a set of edges representing molecular interactions between genes in the pathway $p$. Gene expression profile from RNA-seq is mapped to nodes $V^p$ that are represented as a vector $X_i^p = (x_{i,1}^p, x_{i,2}^p, ..., x_{i,m_p}^p)$, where $i$ is an index for each patient and $m_p$ is the number of genes in the pathway $p$ ($m_p = |V^p|$).

To capture localized gene expression patterns in $G^p$, a spectral convolutional approach is applied on the Laplacian matrix $L^p = D^p - A^p$ of a graph (Bruna *et al.*, 2013; Defferrard *et al.*, 2016). Here, $D^p$ is a weighted degree matrix of $G^p$ and $A^p$ is an adjacency matrix of $G^p$. Based on an eigenvalue decomposition of a graph Laplacian matrix $L^p = U\Lambda U^T|_p$, the spectral convolutional operator is defined as

$$L^p_{spectral} = U g_\theta(\Lambda) U^T X|_p \tag{3.1}$$

where $g_\theta(\Lambda)|_p$ is a polynomial convolution filter that is applied on the diagonal matrix $\Lambda^p$.

$$g_\theta(\Lambda)|_p = \sum_{k=0}^{K-1} \theta_k \Lambda^k|_p \tag{3.2}$$

$g_\theta(\Lambda)|_p$ is represented as a $K$-order polynomial function that works as a convolution filters reaching $K$-hop neighbors. This way, the spectral convolutional operator (Equation 3.1) can capture localized expression patterns in $K$-hop neighbor nodes in a graph. Despite this advantage, it is difficult to use a polynomial convolution filter as is since it takes $O(n^2)$ time to calculate the polynomial filter. In a recent study (Hammond *et al.*, 2011), an approximated polynomial function called Chebyshev expansion is proposed. Using

the Chebyshev polynomial approximation, the spectral convolutional filter is re-defined as

$$L^p_{spectral} \approx U[\sum_{k=0}^{K-1} \theta'_k T_k(\tilde{\Lambda})]U^T X|_p \tag{3.3}$$

where $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0 = 1$ and $T_1 = x$, and $\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I_{m_p}$.

The above filter is used as a convolutional filter. Then, extracted patterns are pooled with neighboring nodes using the Graclus algorithm. Empirically, it is found that max pooling performed better than average pooling for the pathway models, thus a max pooling is used for reducing genes. After convolution and pooling, gene expression profiles are dimensionally reduced into pathway level vectors. In turn, these vectors are given to a MLP and then they concerted to subtype-wise probability vectors.

The entire structure of GCN pathway models is shown in Figure 3.2. To deal with high-dimension low sample characteristics, over 20,000 dimensions of genes and typically less than 1,000 samples, of transcriptome data, dropout and shallow networks for GCN and MLP are used to avoid over-fitting. Cross-entropy loss is used as a cost function.

### 3.2.2 Re-producing comprehensive biological process by Multi-Attention based Ensemble

As described in Section 3.2.1, a GCN pathway model is built for each pathway. Using these GCN pathway models, gene expression profile $X_i$ for each patient is converted into $P$ number of encoded vectors $h^p(X_i)$ (as shown in Figure 3.2) , where $p = 1, 2, ..., P$ (= total number of pathways). To combine encoded vectors of hundred pathways, attention mechanism is used. Each encoded vector $h^p(X_i) \in \mathbb{R}^d$, where $d$ corresponds to the number of cancer subtypes, is concatenated, resulting in a large matrix form $h(X_i) \in \mathbb{R}^{P \times d}$. Attention scores are calculated on the concatenated matrix $h(X_i)$ and attention-based combination

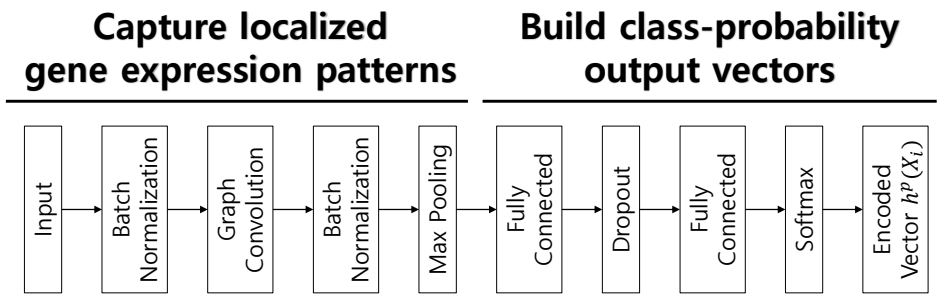



**Figure 3.2:** Structure of GCN pathway model. It consists of a graph convolutional network that extracts gene expression information using pathway topology and a multilayer perceptron (MLP) that encodes the information. To prevent over-fitting, shallow layers with batch normalization and dropout are adopted.

of pathway vectors $\tilde{h}(X_i)$ is generated below as in Equation 3.4.

$$\begin{aligned} W \in \mathbb{R}^{d\times a}, b \in \mathbb{R}^a, u \in \mathbb{R}^a \\ Y = tanh(h(X_i)W + b) \in \mathbb{R}^{P\times a} \\ \alpha = Softmax(Yu) \in \mathbb{R}^P \\ \tilde{h}(X_i)_j = \sum_{k=1}^{P} h^k(X_i)_j \alpha_k \\ \text{where } \tilde{h}(X_i) \in \mathbb{R}^d \end{aligned} \tag{3.4}$$

Our multi-attention based ensemble (MAE) model operates at two hierarchical levels (Figure 3.1(c)): pathway-level attention and ensemble-level attention. The basic mechanism for the pathway-level attention is the same as in Equation 3.4, but multiple attention mechanisms are used to capture various combinations of pathway encoded vectors. Each attention mechanism generates $\tilde{h}(X_i)|_l \in \mathbb{R}^d$, where $l$ is $l$-th attention mechanism. As an ensemble-level attention, multiple pathway-level attention encoded vectors are concatenated, resulting in a form of $\tilde{h}_{MGD}(X_i) = (\tilde{h}(X_i)|_1; \tilde{h}(X_i)|_2; ...; \tilde{h}(X_i)|_L)$

$\in \mathbb{R}^{L \times d}$, where $L$ is the number of pathway-level attention. Then, as in Equation 3.4, an ensemble-level attention vector $\tilde{h}_{fin}(X_i) \in \mathbb{R}^d$ is computed. After the MAE step, $\tilde{h}_{fin}(X_i)$ is used as input to two-layer fully connected MLP for the cancer subtype classification. Cross-entropy loss is used as an objective function.

### 3.2.3 Linking pathways and transcription factors by network propagation with permutation-based normalization

My approach of combining hundred pathways using multi-attention models does provide some insights on how pathways interact differentially among cancer subtypes. Investigation on the difference in gene interaction among subtypes is much more complicated because the number of genes is almost two orders of magnitude larger than the number of pathways. Here, I propose an effective approach of investigating gene interactions using transcriptome data by linking pathways and TFs using network propagation with permutation-based normalization.

Network propagation is typically done by performing random walks on a network. A random work starts with seed nodes that are pre-selected and the seeds have certain amount of information to be propagated. However, performing a random walk on a long path will dilute the information too much, especially when hub nodes with many edges are involved. To avoid this dilution problem, a random walk with restart algorithm (Köhler *et al.*, 2008) is used. The random walk with restart is calculated as below:

$$p^{(t+1)} = (1 - r)Wp^{(t)} + rp^{(0)} \tag{3.5}$$

where $W$ is a column-wise normalized weighted adjacency matrix of a network and $p^{(t)}$ is a vector that contains the propagated values of each node at time step $t$. The seed vector $p^{(0)}$ is a normalized vector of initial values and $r$ is a restart parameter.

When performing the network propagation, constructing network topology and selecting seeds are two most important issues. In the case of the network topology, biological prior knowledge and gene expression data are utilized (Figure 3.3(a)). Based on a PPI from BIOGRID database (Stark *et al.*, 2006), an absolute value of Pearson's correlation is mapped on each edge in the network. To link pathways and the weighted PPI, pathway nodes are added in the network. Edges between a pathway node and genes in the pathway are also added with constant weight 1. On the pathway-PPI network, seed nodes are selected as the pathway nodes and values are assigned by attention weights from the GCN pathway models with multi-attention.

As a result of network propagation, all nodes in a network has propagated values. The propagated values are determined by not only the initial values of seed nodes but also the topology of a network. For example, if a node has a high degree, it may have a larger propagated value than other nodes regardless of seed values. To address this problem, a null distribution of propagated values is computed by a permutation based approach (Figure 3.3(b)). Given pathway attention weights of patient samples, a pathway attention weight is randomly selected from the samples on each pathway and a new random patient is generated. By repeating this procedure 1,000 times and performing network propagation on the random patient samples, a permutation-based network propagation values are generated. From the permutation result, each node in the network is ranked in terms of propagated values and a mean permutation rank is computed by averaging the ranks of all random patients. On a real patient, each node is transformed into ranks which are normalized by the mean permutation rank. Remember that the goal in this step is linking highlighted pathways and TFs. TFs that were curated in the literature (Lambert *et al.*, 2018) are used and ranked as a result of network propagation and normalization.

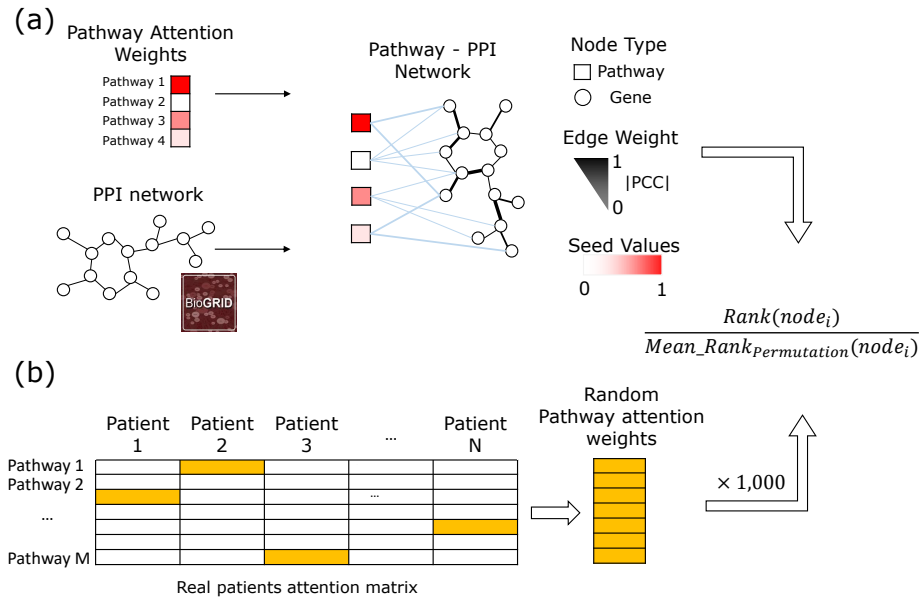



**Figure 3.3:** Construction of pathway-PPI network and Permutation-based normalization. (a) Using pathways and PPI network from BIOGRID, a pathway-PPI network is constructed. A single pathway node is connected with multiple genes that belong to the pathway with a constant weight 1. In addition, to reflect transcriptome profiles of given data, absolute values of Pearson's correlation are mapped on the edges of the PPI network. (b) To address the property that network propagation is vulnerable to high degree nodes, permutation-based normalization is designed. From real patient attention matrix, a random pathway attention weights vector is extracted. By repeating 1,000 times, a set of random patients is generated and network propagation is performed on the data. In a real patient, a rank of node $i$ is normalized by average rank of node $i$ on the random patients.

## 3.3 Results

### 3.3.1 Pathway database and cancer data set

As pathway information, the KEGG pathway database is used(Kanehisa and Goto, 2000). Since the goal is to construct a pathway-based model of transcriptome, some pathways were excluded that are not directly relevant. Specifically, pathways related to drug development were removed. Pathways that are physical clusters of genes were also excluded from the analysis. In addition, pathways with less than five genes were excluded since deep learning models of these small pathways are not feasible. As a result, 287 pathways were used for modeling cancer subtypes. In total, 5,515 genes were included in the 287 pathways, thus the 5,515 genes were used for our analysis. Graph representations of the pathways were extracted using `KEGGgraph` (Zhang and Wiemann, 2009) library in R.

The Cancer Genome Atlas (TCGA) RNA-seq data sets were used as gene expression profiles. RNA-seq data sets were downloaded from Firebrowse (http://firebrowse.org/). Five cancers with subtypes defined, BLCA, BRCA, COAD, PRAD, and STAD, were analyzed. Subtype information for these cancer data sets except BRCA was from the original research papers of each cancer type. In the case of BRCA, the original article (Network *et al.*, 2012) classified subtypes based on microarray data and the number of samples were small since it was one of the early TCGA papers. Thus, BRCA subtypes were re-generated by the PAM50 classification method (Parker *et al.*, 2009) on log2-transformed RNA-seq data. Detailed information about cancer data sets are described in Table 3.1 .

**Table 3.1:** List of cancer data set with subtypes. The number assigned to each subtype represents the number of samples in that subtype. Subtypes with few assigned samples have been removed from the analysis

| **Cancer** | **Total** | **Subtypes** | **Source** |
|---|---|---|---|
| BLCA | 408 | Basal_squamous (142), Luminal (26), Luminal_infiltrated (78), Luminal_papillary (142), Neuronal | * |
| BRCA | 1097 | Basal (230), Her2 (161), LumA (318), LumB (298), Normal-like (90) | ** |
| COAD | 245 | CMS1 (39), CMS2 (78), CMS3 (37), CMS4 (68), NOLBL (23) | * |
| PRAD | 317 | ERG (152), ETV1 (28), ETV4 (14), SPOP (37), other (86) | * |
| STAD | 277 | CIN (138), EBV (25), GS (54), MSI (60) | * |

BLCA: Bladder Urothelial Carcinoma, BRCA: Breast invasive carcinoma, COAD: Colorectal adenocarcinoma, PRAD: Prostate adenocarcinoma, STAD: Stomach adenocarcinoma

* Sources of data set: BLCA (Robertsonet al.(2017)), COAD (Guinneyet al.(2015)),PRAD (Abeshouseet al.(2015)), STAD (Networket al.(2014))

** The subtypes of breast cancer samples were classified using RNA-seq data and
PAM50 as mentioned in the Section 2.4 in the main script.

### 3.3.2 Evaluation of individual GCN pathway models

Before constructing one unified model of transcriptome data, each pathway was modeled as a component. Experiments were performed to see how well each GCN model classified cancer subtypes. Hyper-parameters of each GCN model were determined in 3-fold cross-validation (CV) within training data. Classification performances were measured in terms of weighted F1 score with 10-fold CV (Table 3.2).

The average classification accuracies were 76.39% for BLCA, 66.91% for BRCA, 71.54% for COAD, 70.12% for PRAD, and 78.13% for STAD. There was a huge variation of performances. For example, in the case of BLCA, the maximum value was 90.98% whereas the worst-case value was 46.78%. Because these pathways reflect only a small part of the biological process, some of these pathways were highly correlated with cancer subtypes, but some did not. However, no pathways were commonly singled out in achieving the best performances in all five cancers. Thus, the goal of combining all pathways in one model is well supported by these experiments.

### 3.3.3 Performance of ensemble of GCN pathway models with multi-attention

#### Effectiveness of the multi-attention based ensemble model of GCN pathway models

The first performance evaluation is to see how accurate it can be and how much performance gain can be achieved by combining all pathway models into one model with multi-attention. Hyper-parameters for the MAE model were determined in the same manner as described in Section 3.3.2.

**Performance gain of attention mechanism:** By combining all GCN pathway models to single GCN+MAE models, the performance gain was sig-

**Table 3.2:** Statistics of GCN pathway model performance by the 10-fold cross validation. For all pathways used as input, weighted f1 scores were computed and summarized over 10-fold split results. The maximum, minimum, mean, and standard deviation values were summarized to give a brief overview of the total results

| Cancer | Maximum | Minimum | Average (Std) |
|---|---|---|---|
| BLCA | 90.98 | 46.97 | 76.39 (±8.89) |
| BRCA | 82.72 | 41.32 | 66.91 (±9.10) |
| COAD | 82.79 | 45.16 | 71.54 (±7.89) |
| PRAD | 86.13 | 45.14 | 70.12 (±8.54) |
| STAD | 90.79 | 51.64 | 78.13 (±8.23) |

nificant (Table 3.3). In COAD data, the GCN+MAE model with 11 attentions showed the best F1 score of 87.01% with 4.22% improvement over the best of single GCN pathway model. The other cancer data except STAD were also achieved over 2.7% improvements. Besides, ensemble of multi-attentions notably affected the performance gain. Most of the GCN+MAE models were showed over 2.0% performance gain than single attention without an ensemble level attention (GCN+Single Att). Even with the GCN+Single Att models, the performances were also better than single GCN pathway model. The performance gains were smaller than 1.0% in three cancers, but over 2.0% performance gains were also observed in the other cancers (BRCA and COAD). Thus, these experiments show the effectiveness of attention mechanisms that combine hundred pathway models.

**Comparison with existing methods:** The GCN+MAE model was compared with other classification methods: SAS (Lim *et al.*, 2016), the pathway

**Table 3.3:** Performance comparison of models. The proposed model (GCN+MAE) was compared with other models including the GCN pathway model. The ensemble models using attention mechanism showed better performance than the other classifiers. "GCN + MAE (best)" indicated how many attention mechanisms were used in parentheses. In the parentheses of "GCN best", the ID of the pathway showing the performance was described

| | BLCA | BRCA | COAD | PRAD | STAD |
|---|---|---|---|---|---|
| GCN+MAE (best) | **93.74** (9-Att) | **85.52** (14-Att) | **87.01** (11-Att) | **89.62** (9-Att) | **91.49** (7-Att) |
| GCN+MAE (#class-Att) | 93.48 | 85.22 | 86.25 | 88.52 | 90.8 |
| GCN+ *Single Att | 91.08 | 85.03 | 84.97 | 86.55 | 90.96 |
| GCN best | 90.98 (hsa04151) | 82.72 (hsa05206) | 82.79 (hsa04151) | 86.13 (hsa05200) | 90.79 (hsa04151) |
| †SAS+SVM | 81.51 | 74.41 | 77.54 | 79.25 | 76.08 |
| SAS+RF | 79.12 | 73.54 | 69.44 | 67.02 | 67.00 |
| SAS+MLP | 83.27 | 48.51 | 76.40 | 77.52 | 76.82 |
| ‡RAW+SVM | 89.18 | 82.62 | 78.41 | 82.58 | 86.39 |
| RAW+RF | 79.83 | 77.11 | 74.69 | 68.36 | 76.17 |

* Instead of multi-attention, the GCN pathway models are combined with single attention mechanism

† The pathway activity inference tool from (Lim *et al.*, 2016)

‡ 20,531 genes are used as input features

hsa04151: PI3K-Akt singnaling pathway

hsa05206: MicroRNAs in cancer

hsa05200: Pathways in cancer

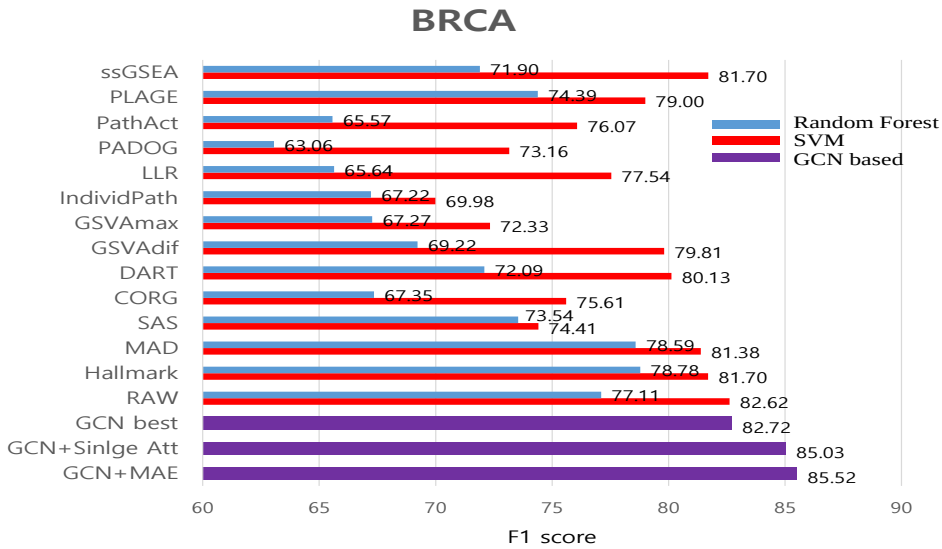



**Figure 3.4:** Cancer subtype classification performance comparison on BRCA data. Including my GCN+MAE model, classification performance tests were performed on 11 pathway activity inference tools and three gene-level based features (all genes, cancer hallmark genes, or top median absolute deviation genes). The GCN+MAE model outperformed other methods. Interestingly, gene-level classifiers showed good classification performance than classifiers using pathway activity inference tools.

activity inference tool, and RAW that used all available 20,531 gene expression. The results are summarized in Table 3.3. A recent study (Lim *et al.*, 2018) compared 13 pathway activity inference tools and SAS showed reasonably good performances on various tests including cancer subtype classification. Thus, SAS was chosen for performance comparison. As pointed out in the study (Lim *et al.*, 2018), a large portion, about 2/3, of gene expression information is lost. Thus, RAW were also chosen to compare classification performances when "all" genes are used. To classify cancer subtypes with SAS and RAW, support vector machine (SVM) and random forest (RF) were used

as classifiers. In addition, pathway activities measured by SAS were trained by MLP classifier. In the experiments of comparing GCN+MAE model with SAS and RAW classifiers, the GCN+MAE model performed significantly better, at least above 2.9%, than both SAS and RAW classifiers. For example, in the case of COAD, SAS+SVM achieved best performance at 77.54%, which was significantly lower than the GCN model (82.79%) and the GCN+MAE model (87.01%). In fact, performance of the SAS classifier was worse than the RAW classifier, which showed that information loss of genes in pathways is substantial. To further explore the relationships between the pathway activity inference tools and gene-level gene sets, I tested 10 additional pathway inference tools and another gene set of cancer hallmarks (Hanahan and Weinberg (2011)) (Figure 3.4; other cancers are not shown). Interestingly, in these experiments, the gene-level classification models performed better than the pathway activity inference tool based models. These experimental results suggest that proper modeling and aggregation of pathway information is important for the pathway-based modeling of transcriptome data.

**Highlighted pathways in breast cancer:** Until now, the usefulness of GCN+MAE model was analyzed in terms of classification performances. Furthermore, it was also investigated that how pathway attention weights were different across subtypes(Figure 3.5 for BRCA data). Pathway attention weights for each patient were determined by a weighted sum of pathway-level attention vectors, and these weights were extracted from the ensemble-level attention vector. For BRCA, the GCN+MAE model was able to highlight pathways that are known to be important in breast cancer. For example, the highlighted pathways were PI3k-Akt signaling (hsa04151) (Paplomata and O'Regan, 2014) and MAPK signaling (hsa04010) pathways (Santen *et al.*, 2002). Overall, patients were well clustered in the heatmap of attention weights in Figure 3.5.

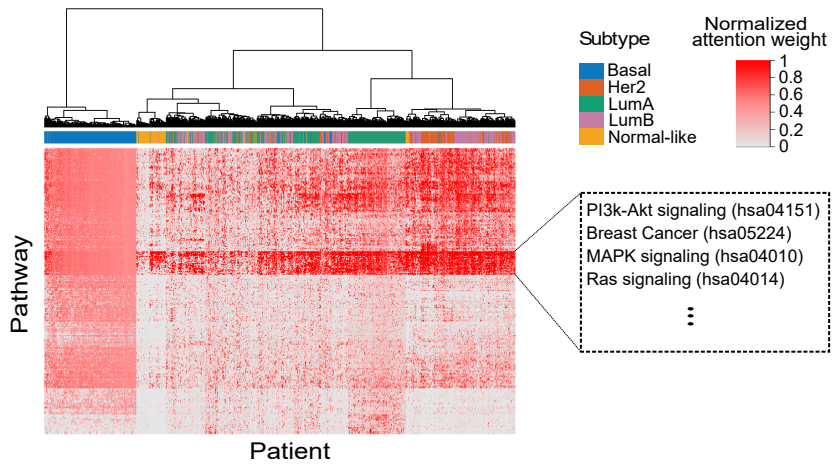



**Figure 3.5:** Heatmap of the attention weight of GCN+MAE model on BRCA data. On the BRCA data, pathway attention weights of each patient were extracted from the best GCN+MAE model in Table 3.3. To better visualization, the whole attention weights were divided at 90th percentile, and values of higher than 1 were forced to 1. Patients and pathways were then reorganized by similarity measure using the manhattan distance and ward D.2 clustering.

In particular, patients of Basal subtype formed a distinct cluster, which could explain aggressiveness of Basal subtype breast cancer in terms of dysregulated or over-activated pathways.

### Effects of the number of multi-attention mechanism

Note that multi-attention were used, thus the number of attentions used to combine pathways would result in performance differences. Thus, I investigated how the classification performance varied concerning the number of attentions. As shown in Figure 3.6 for BRCA data, multiple attention based ensemble models outperformed the single GCN pathway model for all cancer data sets, except STAD cancer data. Performance differences, when different numbers of attentions were used, were less than 1% in most cases.

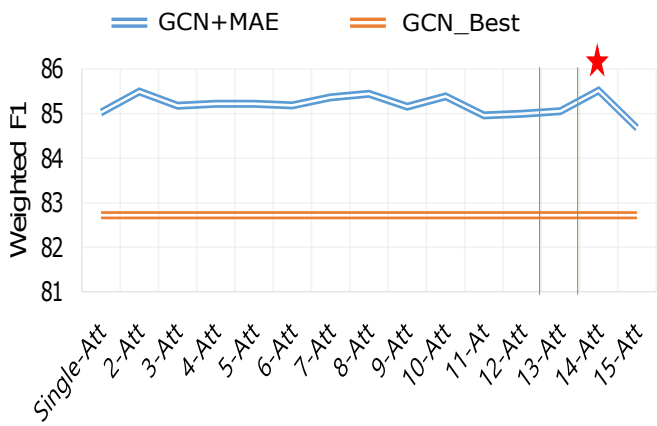
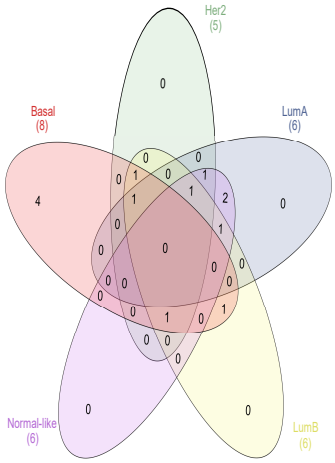



**Figure 3.6:** Performance of GCN+MAE according to the number of attention mechanisms on BRCA. The x-axis of the figure represented the number of attentions from single attention to 15 attentions. The y-axis represented weighted f1 score. The top line was for GCN+MAE and the bottom-straight line was for a single GCN pathway model's best result. Red stars indicated the best classification result point. All results in the figure were calculated in 10-fold cross-validation tests.

Another experiment was clustering analysis of patients. Interestingly, the optimal number of attentions was quite similar to the number of patient clusters. Since input to the GCN+MAE model was a combination of outputs of GCN pathway models, outputs of GCN pathway models were concatenated into a single vector to represent a patient. Then, X-Means clustering was performed on the concatenated vectors. The number of clusters was quite similar to the number of attentions, which could be an explanation of why the GCN+MAE model achieved good performances in subtype classification.




| GO Biological Process | Basal (8) | Her2 (5) | LumA (6) | LumB (6) | Normal-like (6) |
|---|---|---|---|---|---|
| positive regulation of cell proliferation (GO:0008284) | ✓ | | | | |
| negative regulation of apoptotic process (GO:0043066) | ✓ | | | | |
| inflammatory response (GO:0006954) | ✓ | | | | |
| positive regulation of cytokine production (GO:0001819) | ✓ | | | | |
| cytokine-mediated signaling pathway (GO:0019221) | ✓ | | | ✓ | |
| cellular response to cytokine stimulus (GO:0071345) | ✓ | ✓ | | ✓ | |
| regulation of cell proliferation (GO:0042127) | ✓ | ✓ | | ✓ | ✓ |
| regulation of apoptotic process (GO:0042981) | ✓ | ✓ | ✓ | ✓ | |
| regulation of transcription from RNA polymerase II promoter (GO:0006357) | | ✓ | ✓ | | ✓ |
| positive regulation of transcription, DNA-templated (GO:0045893) | | ✓ | ✓ | ✓ | ✓ |
| positive regulation of nucleic acid-templated transcription (GO:1903508) | | | ✓ | | ✓ |
| positive regulation of gene expression (GO:0010628) | | | ✓ | | ✓ |
| negative regulation of transcription, DNA-templated (GO:0045892) | | | ✓ | ✓ | ✓ |

**Figure 3.7:** GO biological processes (BP) enriched in each subtype of BRCA. To find biological functions of top-ranked TFs by network propagation with permutation-based normalization, GO enrichment tests were performed using target genes of different number of those TFs from 5 to 25. Among those tests, consistently detected GO_BP were illustrated as Venn diagram and listed as tables. The number of detected GO_BP were denoted in the parentheses. Similar to the clinical prognosis, almost same GO_BPs were observed in LumA and normal-like subtypes, and the other three subtypes showed also similar results. Unlike other subtypes, Basal subtype contained unique four GO_BP terms that were related to aggressiveness and metastasis of cancer.

### 3.3.4 Identification of TFs as regulator of pathways and GO term analysis of TF target genes

**Subtype-specific TFs:** Multi-attention mechanisms produce which pathways are highlight while making prediction of cancer subtypes. Though some of top 25 highlighted pathways were relatively different among cancer subtypes, most of them were overlapped significantly. Thus, highlighted pathways were not successful in explaining differences in biological functions among cancer subtypes. For this reason, further investigation were performed to focus on dif-

ferences of biological functions among subtypes as described in Section 3.2.3. By normalized rank from network propagation analysis with permutation-based normalization, TFs were ranked by the propagation scores and a majority of ranks were common in each of 10-fold cross-validation experiments. While the TFs in the top 25 highlighted pathways were overlapped among subtypes, TFs that were selected by the network propagation were quite distinct among subtypes.

Top 25 TFs in each subtype of cancers were selected based on the network propagation. For example, in the case of BRCA, different TFs were selected in each of subtypes. Nuclear hormone receptor-related TFs such as *NR1D1, NR1H4, RORA, RORB, RORC* were ranked high in Basal subtype. On the other hand, tumor suppressor genes such as *TP53, FOXO4* ranked top in Luminal A subtype. Then, target genes of these top-ranked TFs were determined from a curated database (Han *et al.*, 2017), and then biological functions of these target genes were identified by GO term analysis by `Enrichr` (Kuleshov *et al.*, 2016). Remember that about 2/3 genes, including many TFs, are not included in the KEGG pathway database. The analysis scheme below is an effective way to investigate differences in biological functions of cancer subtypes from subtype-specific pathways:

$$\text{subtype-specific pathways} \rightarrow \text{TFs} \rightarrow \text{biological functions}$$

**Subtype-specific biological functions:** Enriched GO biological processes (GO_BP) of target genes of TFs were analyzed, varying the number of top TFs from 5 to 25 in a step of 5. Then, consistently detected GO_BP terms regardless of the number of TFs were collected as subtype-specific biological functions (Figure 3.7 for BRCA) GO_BPs enriched in each subtype were represented as Venn diagram using `InteractiVenn` (Heberle *et al.*, 2015). As shown in Figure 3.7, GO_BP terms enriched in Basal subtype are most distinct compared

to GO_BP terms enriched in other subtypes, which could be a good explanation on why Basal subtype is most aggressive. On the other hand, LumA and Normal-like subtypes shared almost the same GO_BP terms including positive regulation of nucleic acid-template transcription (GO:1903508) and positive regulation of gene expression (GO:0010628), which also explains better prognosis of LumA compared to other subtypes. In addition, three subtypes, such as Her2, LumB, and Basal, shared cellular response to cytokine stimulus (GO:0071345) that was known as important factors of breast cancer development and metastasis (Esquivel-Velázquez *et al.*, 2015; Eichbaum *et al.*, 2011).

GO_BP terms that were enriched only in Basal subtype could explain its aggressive phenotype. Positive regulation of cell proliferation (GO:0008284) and negative regulation of apoptotic process were closely related to cancer cell proliferation and they were known as one of the hallmarks of cancer (Hanahan and Weinberg, 2011). Basal subtype cancer had a high rate of proliferation than other subtypes. This dysregulated cell proliferation resulted in worse prognosis and treatment options are difficult to choose for patients (Castelvecchi, 2016; Cakir *et al.*, 2012). Two more enriched GO_BP terms were inflammatory response (GO:0006954) and positive regulation of cytokine production (GO:0001819). Cytokines were a family of proteins related to immune systems. Inflammation and escape of immune destruction were also members of hallmarks of cancer (Hanahan and Weinberg, 2011). Many studies reported that Basal subtype breast cancer exhibited differential expression of inflammation related genes and stronger immunogenicity than the other breast cancer subtypes (Liu *et al.*, 2018; Hartman *et al.*, 2013). For these reasons, inflammatory related cytokines were considered to be immunotherapeutic targets of Basal or triple-negative breast cancer (Kim *et al.*, 2015; Fabre *et al.*, 2018).

# Chapter 4

# Detecting sub-modules in biological networks with gene expression by statistical approach and graph convolutional network

In the previous study, I used biological pathways as single units. A biological pathway is to represent a distinct biological process. However, a pathway can contain multiple biological functions. Thus, to investigate biological mechanisms under specific conditions, e.g., cancer, it is sometimes necessary to dissect a pathway into a set of smaller units, each of which can represent a single biological function. To address this issue, an algorithm is needed to determine functionally coherent subgraphs of activated genes.

## 4.1 Motivation

The advent of high throughput technologies for transcriptome profiling, such as microarrays or RNA-Seq, has changed the paradigm of transcriptome analysis from the gene-centric research to the genome wide investigation of a biological

mechanism (Luo *et al.*, 2010; Jin *et al.*, 2014). The most widely used analysis technique is to determine a list of differentially expressed genes (DEGs). This approach can be useful to find genes that play important roles between control and treated conditions, e.g., disease and healthy patients. However, the DEG-based analysis has serious limitations. For example, consider transcriptome data for investigating disease conditions. Disease is not caused by an abnormal activity of a single gene but complex perturbation involving many genes is the cause and result of a disease (Barabási *et al.*, 2011). Investigation on biological mechanisms underlying difference in transcriptomic abundance of a number of genes is challenging (Khatri *et al.*, 2012). To overcome these limitations, the pathway based approach has emerged and routinely used to derive informative biological insights from transcriptome data (Khatri *et al.*, 2012; Luo *et al.*, 2009; Kelder *et al.*, 2010; García-Campos *et al.*, 2015). A biological pathway is a graph representation showing how genes interact based on the literature information and experimental validations. The most well known pathway database is KEGG(Kanehisa and Goto, 2000) and there are several well curated pathway databases such as REACTOME (Croft *et al.*, 2010), NDEx (Pratt *et al.*, 2015), and PANTHER (Mi *et al.*, 2013).

### 4.1.1 Pathway based analysis of transcriptome data

To investigate which pathway is activated and suppressed, gene expression information from transcriptome data should be mapped to nodes or genes of the pathway. Since genes are inter-connected in complex ways, sophisticated bioinformatics methods are needed to investigate activation or suppression status of a pathway. Bioinformatics methods to analyze pathway activation or suppression status can be categorized into three groups: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology (PT) (Khatri *et al.*, 2012). ORA statistically measures how much fraction of genes in the

specific pathway are included in a gene set, e.g., a set of DEGs. Perturbed pathways are selected in terms of the statistical significance that is calculated by Fisher exact test or chi-square test (Zeeberg *et al.*, 2003; Bindea *et al.*, 2009). ORA does not consider the gene expression quantity information by treating all DEGs are equally, thus ORA often fails to characterize differences between phenotypes in terms of gene expression and pathway activation/suppresion.

The second generation pathway analysis method, FCS, is based on the fact that biological mechanisms are affected by not only large changes in few genes but also many functionally related genes with weak transcription level. To aggregate effect of all genes in a pathway, FCS methods defines a score of each gene based on the statistical significance. Then a score of pathway activation and suppression is defined by simply aggregating scores of each gene in the pathway. Statistical significance of the aggregated score is tested against null hypothesis that the pathway gene set is associated with phenotypes no more than the genes not in the pathway (competitive) or the gene set is differentially expressed between phenotypes (self-contained). The most widely used FCS methods are gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005) and Pathifier (Drier *et al.*, 2013). GSEA determines whether a set of genes (e.g., pathways) are statistically different between two phenotypes by measuring whether those genes are randomly distributed or located in the top or bottom of a list of genes (e.g, list of DEG). While GSEA measures the difference between phenotype groups, Pathifier calculates the pathway score individually. Using principal component analysis (PCA) and principal curve that captures the variations in whole sample, gene-level information is changed to a single pathway-level score.

Like ORA methods, a major issue with FCS methods is that these methods do not consider topological information of pathways such as interaction between genes and gene regulation information. To address this issue, a new class

of pathway analysis methods, PT, emerged. The main difference between FCS and PT is use of topology information when gene-level scores are measured. The well known PT methods are SPIA (Tarca *et al.*, 2009) and PARADIGM (Vaske *et al.*, 2010). SPIA determines significant pathways based on two types of evidences. One is a significance score generated by ORA. The other is the perturbation score of pathway by propagating expression changes between two phenotypes via topological structures of pathway. PARADIGM infers activity of a specific pathway in a sample-level using a factor graph that is constructed from pathway. PARADIGM is designed to handle multi omis data (gene expression, methylation, copy number variation) and utilizes a belief propagation algorithm.

There is another major challenge in measuring pathway activities. Pathway is designed to capture series of molecular interactions that change state of a cell or produce certain chemicals. Thus a pathway consists of multiple biological functions, not a single homogenous function. To handle this problem, it is necessary to divide a pathway into multiple sub-pathways each of which has a single biological function. Overbeek et al. (Overbeek *et al.*, 2005) pioneered to use this concept by defining and using subsystems to annotate genomes. Since then, subsytem/subpathway based approaches are used to deduce more accurate and sensitive biological interpretations. Chang, Jeffrey T. et al. (Chang *et al.*, 2009) used an approach to deconstruct a pathway into modules so that each module can have a single molecular function and also can model complex, non-linear relationship among genes in the pathway. In addition, clustering approaches are also used to decompose pathway into functional modules. Barabási et al. (Barabási *et al.*, 2011) used a network clustering method to identify drug target biomarkers considering functional relationship among nearby genes. A recent study (Lim *et al.*, 2016) developed a new method of decomposing a pathway into functional sub-pathways

using clusters obtained from the protein interaction network. This method defined and used an edge activity score that considers explicit gene expression information from RNA-seq and network centrality information of each gene. Recently, a number of different approaches has been developed to measure activity of sub-pathways, rather than a whole pathway. Sub-pathway activity measurement tools are designed to identify activated subpaths between two phenotypes: PATHOME (Nam *et al.*, 2014), TEAK (Judeh *et al.*, 2013), and MinePath (Koumakis *et al.*, 2016).

### 4.1.2 Challenges and Summary of Approach

Although there has been a significant development in measuring pathway or sub-pathway activities over the years, several technical challenges remain to be resolved.

**Challenge 1: Use of explicit gene expression information** Existing pathway methods are designed for microarray data and they do not utilize explicit gene expression information from RNA-seq that is known to produce more accurate gene expression information (Wang *et al.*, 2009). Some methods are designed to handle microarry data only since the analysis method assumes some specific distributions to determine subpaths, e.g., a bayesian network based subpath identification method (Judeh *et al.*, 2013). Most of existing methods convert gene expression into correlation between two genes or binary notation (up-regulated / down-regulated), thus not using explicit gene expression quantity.

**Challenge 2: Measuring activity of subpath consisting of multiple nodes or edges** Determining subpaths that exhibit differential activities in different phenotypes requires to handle multiple genes (nodes) or edges. Mapping gene expression information to the corresponding gene in the pathway

is not very helpful in determining subpaths since no topological information is considered. Some existing methods measures activities of edges but this is mostly done by correlation analysis which is not additive. Not being additive is a serious hurdle in determining subpaths.

**Challenge 3: Multi-class differential subpath activity to determine condition specific subpath activity** Recently, transcriptome data is used to compare multiple, more than two, phenotypes. This trend is expected to continue in an increasing way since transcriptome information from RNA-seq allows us to compare arbitrary number of phenotypes. Traditional approaches using the concept of up/down regulation is not extensible to multiple-class comparisons.

**Challenge 4: Determining differential subpath activity using bulk cell sequencing data** Sequencing requires a good quantity of RNA sufficient for sequencing experiments. Increasingly, RNA are obtained from a bulk of cells that consist of cells of different types. In this case, determining differential subpath activity is even more challenging since extending subpath by adding nodes (genes) or edges requires rigorous criteria.

In this study, a software package, MIDAS, was designed and implemented that considers all of the four issues above. Below I briefly summarized the strategy to address the issues.

1. MIDAS utilized explicit gene expression quantity information from RNA-seq.

2. An edge activation measurement technique (Lim *et al.*, 2016) was extended for determining subpaths with differential activities. See Section 4.2.2 for details.

3. The multi-class issue was considered in a statistical approach. See Section 4.2.3 for details.

4. MIDAS used a greedy subpath extension method with exponentially increasing criteria. See Section 4.2.3 for details.

Although MIDAS considered the four issues mentioned above, there are drawbacks that are not considered. MIDAS overlooked the fact that one gene belongs to several pathways. This is because pathways are the result of partitioning the entire biological system for ease of interpretation. To improve this, I designed to follow up on MIDAS using PPI network which is bigger network than pathway.

A graph convolutional network is used to draw significant features from the PPI network, taking into account gene expression and gene interactions. Graph convolutional operations exist for both spectral and non-spectral approach. In the second study, spectral graph convolution (Defferrard *et al.*, 2016) was used because pathways are small networks. In this study, rather than utilizing Fourier operation and approximation step on huge network, a non-spectral method was used (Kipf and Welling, 2017). When selecting features related

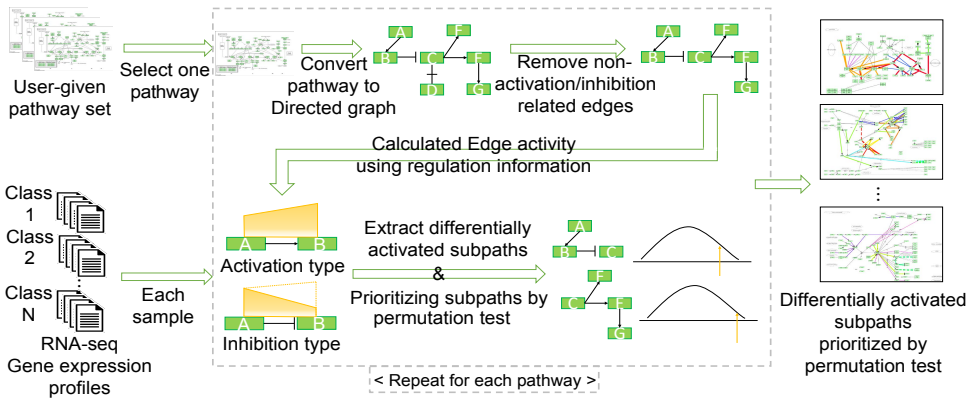



**Figure 4.1:** The workflow of MIDAS. From user-given RNA-seq gene expression profiles and target pathway set, differentially activated subpaths among classes are determined.

to gene interactions with graph convolution, multi-hop convolution operation was performed by considering long range interaction between genes.

Additionally, the class activation map (CAM) approach was used to extract subnetworks that are important for classifying subtypes on networks (Zhou *et al.*, 2016). CAM is a technique used for image classification problems. When classifying a class, it is a technique that indicates what part of the image a model focuses and predicts the class. Using this, it is possible to investigate which subnetwork is closely related to the class on the network. By using the CAM method, 78%  85% of the performance is obtained in the subtype classification problem of breast cancer data, and the subnetwork is extracted according to the subtype.

**[Problem Definition of this study]**

Given a set of pathways $i = 1, 2, ..., m$ or a PPI network and gene expression data $X$ with $N$ patients

$< Input >$

$G_i$ : a graph of pathway of PPI network, $G_i = (V_i, E_i)$

$V_i$ & $E_i$ : a set of genes and interactions in the pathway $G_i$

$X_i$ : a gene expression matrix, $X \in \mathbb{R}^{N \times |V_i|}$

$< Output >$

$Y$ : Cancer subtype of given N patients, $Y = \{0, 1, 2, ..., c\}^N$,

$c$ : number of classes

$< Model >$

Extract subpaths in the pathway $i$ or in the PPI network by statistical or GCN approach

Using the subpaths as features, built a classifier $f$ to predict cancer subtype $Y'$

## 4.2 Methods

The overview of the method is illustrated in Figure 4.1. Pathway information is obtained by KEGG database (Kanehisa and Goto, 2000).

- Input : RNA-seq gene expression data with multi class, target KEGG pathway sets to be analyzed
- Parameters : Start_threshold, Increase_moment, Permutation p-value cut-off
- Output : Differentially activated subpaths prioritized by the permutation test

With an input of RNA-seq and target pathways, differentially activated subpaths are determined in four steps. Each of the target pathways is converted into a directed graph. Then gene expression profiles are mapped into the graph and edge activities are measured. Differentially activated subpaths among classes are constructed by a greedy seed and extension method with the exponential decaying threshold. Finally, subpaths are prioritized by the permutation test. Details of each step are in the following subsections.

### 4.2.1 Convert single KEGG pathway to directed graph

Each target pathway is converted to directed graph with preserving entry information and edge regulation information by a R package, KEGGgraph (Zhang and Wiemann, 2009). Entry in a KEGG pathway may contain several genes having similar biological functions, rather than a single gene. In addition, complex of entries, that work together such as CDKs (Cyclin-dependent kinase) and Cyclins in Cell cycle, are denoted as a group entry. These entries and group entries are used as graph node to reflect curated information from KEGG database (Kanehisa and Goto, 2000). In case of regulation, only activation or inhibition edges are used for further analysis like SPIA (Tarca *et al.*, 2009). For example, "binding" or "dissociation" are excluded during graph construction.

### 4.2.2 Calculate edge activity for each sample

The activation status of edge in the constructed graph is measured for each sample. Biologically, genes interact with nearby genes and activity is determined by considering these interactions rather than by a single transcriptional abundance. To consider topological importance and expression levels of genes, the Lim et al.'s approach was extended (Lim *et al.*, 2016). The method in (Lim *et al.*, 2016) measured edge activity on a undirected protein-protein in-

teraction (PPI) network. In this study, graph edge has two type of regulation information: activation and inhibition. For the activation type edge, the same measurement in (Lim *et al.*, 2016) is used. However, for the inhibition type edge, inhibitory mechanism is reflected in the measurement as below.

- "Activation" type edge A $\rightarrow$ B:
  $Act_e = \frac{1}{2} \times \frac{\{ce(A) \times expr(A) + ce(B) \times expr(B)\}^2}{expr(A) + expr(B)}$
- "Inhibition" type edge A $-|$ B:
  $Act_e = \frac{1}{2} \times \frac{\{ce(A) \times expr(A) + ce(B) \times (max\_expr(B) - expr(B))\}^2}{expr(A) + (max\_expr(B) - expr(B))}$

  where
  - ce(A) : closeness centrality of node A
  - expr(A) : average gene expression of node A (Many genes are included in a single node)
  - max_expr(A) : maximum average gene expression of node A in the whole samples.

ce(A) is closeness centrality of node A in the target pathway graph. Because the pathway graph contains gene interaction information and signaling mechanisms, closeness centrality represents topological importance of the node in the target pathway graph. It is used in calculation of edge activity to give more weights to the gene that regulates several genes in the pathway.

### 4.2.3 Mining differentially activated subpath among classes

The goal is to determine subpaths with different activities across phenotypes. This problem is computationally intensive since it is needed to consider all possible pairs of genes in a pathway. In addition, each candidate subpath should be tested if the subpath has phenotypically different. Thus a greedy seed-and-extension algorithm was designed and implemented. As mentioned

in Introduction, the pathway is composed of several biological processes. Thus determining differentially activated subpath is not an easy task. A reasonable search strategy is needed to explore the huge search space. Thus greedy seed & expansion technique with exponential decaying threshold was implemented . To begin with, subpath activity is defined as an average of activity value of edges belonging to the subpath. Then, distributions of subpath activities are created for each class, and the distribution difference among the classes is measured with statistical test, "kruskal-wallis test". To avoid incorrect extension, the algorithm enforces a very stringent criteria with exponentially decaying threshold values as the subpath gets longer. Default value of Start_threshold is 0.05 and Increase_moment is empirically determined according to class number and sample size. The subpath determination algorithm works as below.

---

[*Input*]: RNA-seq Gene expression Data **D**, Pathway graph **G**, Start_threshold, Increase_moment

1. Calculate edge activity on each edege $e \in$ in **G** per each sample in **D**.
2. Perfom kruskal-wallis test on each edge $e$, i.e., size 1 (= two nodes) subpath.
3. **Seed Selection**: Select the most significant edge in the graph as current subpath **SP**, i.e., Seed. In addition, set the threshold $\delta$ to the Start_threshold value.
4. **Expansion Step-1**: Search neighbor edge set NE of the current subpath **SP** (initially, it is seed). For each edge $e$ in the NE, create a temporary new subpath **TP** by adding $e$ to the current subpath **SP**, and generate a statistical statistic through the kruskal wallis test. Select the edge $e^*$ that produces the best kruskal wallis statistic value.

- Expansion candidate edge $e^* = \arg\max_{e \in \mathbf{NE}} KW(\mathbf{TP}), \mathbf{TP} \leftarrow \mathbf{SP} + e$

5. **Expansion Step-2**: Expand current subpath **SP** by adding the edge $e^*$ selected from Step3 if the significance of the new subpath calculated from the statistical test is less than the given threshold $\delta$ (initially, it is same as Start_threshold).

   - Expanded new subpath $\mathbf{SP} \leftarrow \mathbf{SP} + e^*$, if $p - value(\mathbf{SP} + e^*) < \delta$

6. **Expansion Step-3**: If the expansion is successful, the given threshold $\delta$ becomes tight by exponential decaying. Multiply the threshold by an amount Increase_moment to construct a tight new threshold and return to Step 4. Else, remove the current subpath from the graph.

   - Exponentially decaying new threshold $\delta \leftarrow \delta * \text{Increase\_moment}$

7. Repeat Step 3 to 6 until no edges remain in the graph.

[*Output*] : Differentially activated subpaths

---

### 4.2.4 Prioritizing subpaths by the permutation test

Once, differentially activated subpaths are determined, significance of subpaths is measured by a permutation test to reflect intra and inter pathway relationship. For each pathway, a permutation p-value is calculated by generating a distribution of random subpaths of the same size by creating a null distribution for each size. For example, a subpath of size k, $S_k$, K edges are randomly selected from the pathway and then a random subpath activity is constructed to calculate the Kruskal-Wallis statistics. This operation is repated 10,000 times to create a null distribution for size k. A permuation p-value of $S_k$ is measured by Equation 4.1.

$$Permutation\, p-value(S_k) = \frac{\sum_{s_i^k \in S^k} I(kw(s_i^k) \leq kw(S_k))}{10,000} \tag{4.1}$$

- kw($S_k$)) : Kruskal-Wallis statistic of given size k subpath
- $S^k$ : null distribution of size k subpath

### 4.2.5 Extension: graph convolutional network and class activation map

From the second and third study, GCN+MAE and MIDAS, the usefulness of the pathway is tested. However, pathway databases contain only a small part of entire genes and information of most genes are lost. A protein-protein interaction (PPI) network, rather than pathways, contains almost genes of the living organisms. To utilize a large network information and focus on specific nodes in the network, i.e., important genes, a graph convolutional network and class activation map (Zhou *et al.*, 2016) approach is performed for extension of MIDAS method.

For a given graph $G = (V, E)$, an adjacency matrix $A$ is determined. An operation of graph convolution of an input $X$ and a weight matrix $W$ is followed the widely used graph convolutional framework (Kipf and Welling, 2017) like below.

$$f(X, A) = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}XW) \tag{4.2}$$

with $\hat{A} = A + I$, where $I$ is the identity matrix and $\hat{D}$ is the diagonal node degree matrix of $\hat{A}$.

A single graph convolutional operation by Equation 4.2 aggregates information of first neighbor nodes. On the biological network, multiple genes interact with each other and form a sub-network with second, third, or more neighbor nodes and range of interacting neighbor nodes are different on each

node. To reflect these properties, attention-based aggregation of multi-hop graph convolution is used for the model. As Equation 4.2, a $l+1$-th hop is determined by a result of $l$-th graph convolution layer like below.

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}(H^{(0)} + H^{(l)})W^{(l+1)}) \tag{4.3}$$

with $H^{(0)} = X$ and when $l = 0$, $H^{(l)}$ term is ignored. Convolutional results of $H^{(l)}$, $l = 1, 2, ..., k$ are aggregated by attention mechanism.

$$H_i^{Att} = \sum_{l=1}^{k} \alpha_l H_i^{(l)}$$

where $H_i^{(l)}$ is a feature of node $i$ in $l$-th convolutional layer and $\alpha_l$ is computed by an attention mechanism like Equation 3.4.

A classifier for prediction of breast cancer subtypes is built by class activation map approach. Using a final graph convolutional unit of Equation 4.2, a channel of each node is transported into class number of filters. On each class filter, global average pooling (GAP) is performed and results of GAP are passed into the final softmax layer to predict classes.

## 4.3 Results

To utilize the tool into biological data, breast cancer is selected as test data. Breast cancer is the most prevalent cancer in women worldwide (Stewart *et al.*, 2016). Breast cancer is a disease that has been studied for decades and it is well-classified as clinically important five molecular subtypes (Basal, Her2, Luminal A, Luminal B, Normal-like) by the PAM50 gene set (Parker *et al.*, 2009; Dai *et al.*, 2015). In addition, the tumor is of a heterogeneous cell population, thus breast cancer may be best to show the utility in terms of the biological and clinical significance of subpaths determined by MIDAS.

Normalized gene expression profiles (Level 3) from RNA-seq data of breast invasive carcinoma were downloaded from TCGA Research Network:

**Table 4.1:** Pathway set used in analysis. From KEGG database, 10 pathways related to breast cancer are curated

| KEGG pathway ID | Name | Reference |
|---|---|---|
| hsa04010 | MAPK signaling pathway | (Menendez *et al.*, 2005; Mao *et al.*, 2010) (Mirzoeva *et al.*, 2009) |
| hsa04014 | Ras signaling pathway | (Lo *et al.*, 2004) |
| hsa04110 | Cell cycle | (Biswas *et al.*, 2000; Porter *et al.*, 1997) |
| hsa04151 | PI3K-Akt signaling pathway | (Berns *et al.*, 2007; Tokunaga *et al.*, 2006) |
| hsa04210 | Apoptosis | (Abedin *et al.*, 2007) |
| hsa04310 | Wnt signaling pathway | (Li *et al.*, 2003; Howe and Brown, 2004) (Schlange *et al.*, 2007; Katoh and Katoh, 2007) |
| hsa04390 | Hippo signaling pathway | (Chen *et al.*, 2012; Lai *et al.*, 2011) |
| hsa04550 | Signaling pathways regulating pluripotency of stem cells | (Wang *et al.*, 2011; Katoh and Katoh, 2007) (Hennessy *et al.*, 2009; Yang *et al.*, 2013) |
| hsa04668 | TNF signaling pathway | (Stuelten *et al.*, 2005) |
| hsa04915 | Estrogen signaling pathway | (Osborne *et al.*, 2005; Thomas *et al.*, 2005) (Massarweh *et al.*, 2008) |

`http://cancergenome.nih.gov/` and ten breast related pathways were selected as shown in Table 4.1. The parameter values were set as follows (Start_threshold: 0.05, Increase_moment: 1e-15, Permutation p-value cut-off: 0.1). Graphical images of pathway graph is generated using KEGGParser (Nersisyan *et al.*, 2014) in Cytoscape (Shannon *et al.*, 2003)

The utility of MIDAS was demonstrated in four ways. The 36 subtype specific subpaths that MIDAS are well supported in the literature in Section 4.3.1.Subsequently, these subpaths have a good discriminant power for cancer subtype classification in Section 4.3.2 and also have a prognostic power in

terms of survival analysis in Section 4.3.3. Finally, performances of MIDAS are compared with a recent subpath prediction method, PATHOME (Nam *et al.*, 2014) in Section 4.3.4.

### 4.3.1 Identifying 36 subtype specific subpaths in breast cancer

From breast cancer gene expression data with five subtypes and 10 target KEGG pathways, 36 subpaths were determined as summarized in Table 4.2. Apoptosis (8 subpaths), PI3K-Akt signaling pathway (6 subpaths), cell cycle (4 subpaths), MAPK signaling pathway (4 subpaths), RAS signaling pathway (3 subpaths), TNF RAS signaling pathway (3 subpaths) showed significant subtype specific pathway activities. In cancer vs. normal, differential activities in these pathways is obvious but identifying pathway activity differences in breast cancer subtypes is not trivial. In terms of subpath length, the longest one was of 14 genes and the average length was about 6 genes (6.19). Average subpath activities among breast subtypes are illustrated in Figure 4.2(a). The ranks left outside the heatmap are subpath ranks and the color map right outside the heatmap indicates which pathway each subpath are derived from. Subpath activities were prominent in aggressive basal subtype samples and in normal samples, which is quite intuitive since 10 target pathways were selected based on the relevance to cancer. Most important information from this RNA-seq analysis is that MIDAS were successful in determining subpaths that have distinct activities in five subtypes. In this section, subpath activities in two pathways, cell cycle and apoptosis, are discussed. Figure 4.2(b) and (c) shows subpath activities in Apoptosis (hsa04210) and Cell cycle (hsa04110), respectively. In Apoptosis (hsa04210), eight subpaths were differentially activated from rank4 (highest rank) to rank31 (lowest rank). Those subpaths are associated with caspase related regulation process (Kumar, 2007), pro/anti-apoptotic function induced by BCL2-family (Czabotar *et al.*,

**Table 4.2:** Pathway Membership & Size information about significant subpaths. (a) contains how many subpaths are extracted from specific pathway. (b) contains occurrence information of significant subpath with certain number of nodes. Pathway information is described here: hsa04010/MAPK signaling pathway, hsa04014/Ras signaling pathway, hsa04110/Cell cycle, hsa04151/PI3K-Akt signaling pathway, hsa04210/Apoptosis, hsa04310/Wnt signaling pathway, hsa04390/Hippo signaling pathway. hsa04550/Signaling pathways regulating pluripotency of stem cells, hsa04668/TNF signaling pathway, hsa04915/Estrogen signaling pathway

**(a)** Pathway Membership information

| Pathway | # of Subpaths |
|---|---|
| MAPK signaling pathway | 4 |
| Ras signaling pathway | 3 |
| Cell cycle | 4 |
| PI3K-Akt signaling pathway | 6 |
| Apoptosis | 8 |
| Wnt signaling pathway | 4 |
| Hippo signaling pathway | 1 |
| Signaling pathways regulating pluripotency of stem cells | 1 |
| TNF signaling pathway | 3 |
| Estrogen signaling pathway | 2 |

**(b)** Subpaths size information

| # of Nodes | # of Subpaths |
|---|---|
| 1 | 4 |
| 2 | 1 |
| 3 | 1 |
| 5 | 6 |
| 6 | 10 |
| 7 | 4 |
| 8 | 4 |
| 9 | 2 |
| 10 | 1 |
| 11 | 2 |
| 14 | 1 |

2014), pro-apoptotic genes like TP53 and FAS (Papaliagkas *et al.*, 2007), and so on. In the meanwhile, in case of Cell cycle (hsa04110), four subpaths were differentially activated from rank 1 (highest rank) to rank29 (lowest rank). Most of those subpaths are related with Cyclin and CDK complexes that are important to regulate cell cycle phase transition such as G1/S phase or G2/M phase (Keyomarsi *et al.*, 2002; Michalides *et al.*, 2002; Casimiro *et al.*, 2012). Mitosis related metaphase/anaphase transition process is also differentially activated (Bharadwaj and Yu, 2004). In summary, the subpaths of the two pathways were successful in explaining subtypes that were different in terms of cell growth and cell death.

### 4.3.2 Subpath activities have a good discrimination power for cancer subtype classification

In breast cancer, the molecular subtype classification is important because the subtypes have different disease characteristics and clinical outcomes (Dai *et al.*, 2015; Sotiriou *et al.*, 2003; Reis-Filho and Pusztai, 2011). To test the classification power of differentially activated subpaths, a random forest algorithm is used in a 10-fold cross validation scheme. Expression profiles were divided into 10 subsets while preserving the subtype ratio. Differentially activated subpaths were determined using the train data and activities of these subpaths were used as features to generate a random forest classification model which was used to predict subtypes of samples in the test data. In the 10-fold cross validation test, the average classification accuracy was 78.41%. To compare the predictive power of subpaths, another 10-fold cross validation using random forest was performed using all genes ($>$ 20,000 gene). In this case, the average classification accuracy was 79.44%. While the accuracies of two classification tests were similar, the number of genes used for classification was very different. Differentially activated subpaths contained only 221 entries and 478

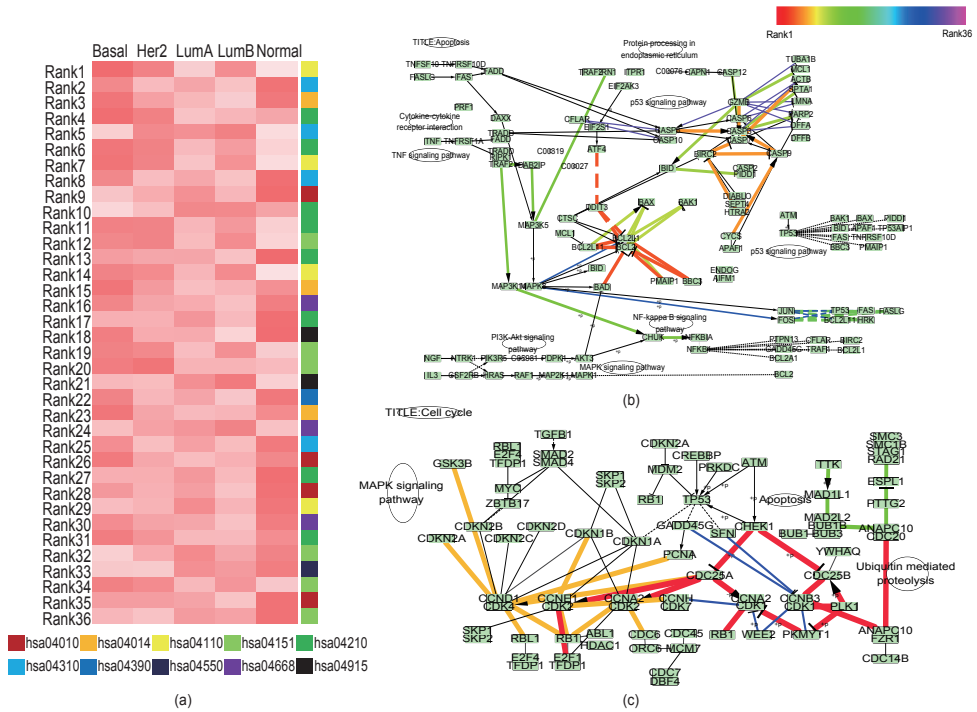



**Figure 4.2:** Average subpath activity among breast cancer subtypes and Subpaths result. (a) average subpath activity is coded as color heatmap. Red color denotes higher subpath activity and white denotes lower subpath activity. (b) and (c) are results where differentially activated subpaths are located. Those subpaths are decoded as rainbow color scheme and edge widths according to their rank. The higher rank subpath is more thicker and red side color. (b) is result of Apoptosis (hsa04210). (c) is result of Cell cycle (hsa04110).

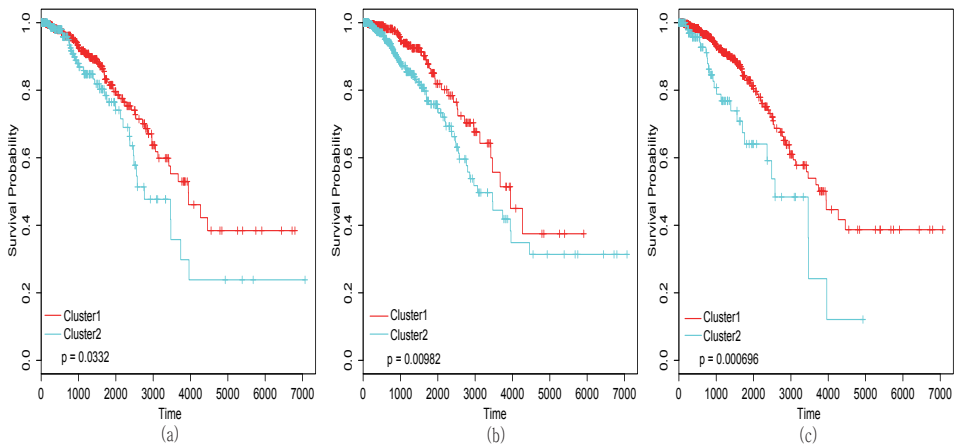



**Figure 4.3:** Survival analysis using differentially activated subpaths with different clustering algorithms. Using differentially activated subpaths as features and different clustering algorithms, clustering and survival analysis are performed 10times for each algorithms. Median p-value result is used as representative result. (a)-(c) show survival analysis results of different clustering algorithms. (a) K-means clustering. (b) robust sparse k-means clustering (RSKC). (c) hierarchical clustering. All of analysis results show statistically significant difference between two groups.

genes (about 2.3%). This shows that the algorithm, MIDAS, was successful in selecting a small number of core genes that can be used to explain cancer subtypes, without sacrificing classification accuracies.

### 4.3.3 Subpath activities have a good prognostic power for survival outcomes

Another experiment in terms of prognostic power was performed to show the utility of the subpaths. To predict survival outcome, it is necessary to divide samples into distinct groups. Using subpath activities, all samples were

divided into two groups. Three different clustering algorithms were used to eliminate the potential bias by different clustering methods. Three different clustering methods were a standard k-means clustering algorithm, a robust k-means clustering (RSKC), and a hierarchical clustering algorithm. RSKC is a variation of the k-means clustering that is designed for data that have noise variables or outlier samples (Kondo *et al.*, 2016). The hierarchical clustering were performed with the euclidean distance and the ward.D2 agglomeration method. To divide samples into good/bad prognostic groups, all of the clustering methods produced two representative clusters. P-value is calculated by a log-rank test. Clustering results may be different due to the k-means seed or the agglomeration order, so all tests were repeated 10 times and the median p-values were shown in Figure 4.3. All survival analysis results were statistically significant at the level of p-value of 0.05: k-means clustering (p=0.032), RSKC (p=0.00982), hierarchical clustering (p=0.000696). Table 4.3 summarizes ratios of samples in each subtypes in two clusters. In case of Cluster1 (the good prognostic group), more samples in less aggressive breast cancer subtypes were included than Cluster2 (the bad prognostic group). Especially, in the k-means and the hierarchical clustering results, over 90% of samples in Cluster1 were those in less aggressive breast cancer subtypes. In the meanwhile, RSKC showed a good prognostic result probably because samples with aggressive breast cancer subtypes, e.g., basal, and normal samples, were divided well into one of the two clusters.

### 4.3.4 Comparison with an existing tool, PATHOME

Performance of MIDAS was compared with PATHOME (Nam *et al.*, 2014). PATHOME is a method for detecting differentially expressed subpaths from KEGG pathway. PATHOME uses template subpaths that are generated by DFS (Depth First Search) from a start node (ex. genes in the membrane) to

**Table 4.3:** The rate at which the subtype is divided into clusters from Survival analysis. This is the result of summarizing how many percent of each subtype is included into clusters obtained from survival analysis with different clustering algorithms. Cluster1 is good-prognostic cluster and Cluster2 is bad-prognostic cluster. (a) K-means clustering, (b) Robust sparse K-means clustering (RSKC) (c) hierarchical clustering. In the Cluster1, less aggressive subtypes such as LumA and Normal are included more than Cluster2. In addition, the ratio of aggressive subtypes belonging to Cluster2 is higher than that of less aggressive subtypes

**(a)** k-means clustering

| Subtype | Cluster1 | Cluster2 |
|---|---|---|
| Basal | 54.63% | 45.37% |
| Her2 | 51.28% | 48.72% |
| LumA | 91.05% | 8.95% |
| LumB | 59.86% | 40.14% |
| Normal | 95.51% | 4.49% |

**(b)** RSKC

| Subtype | Cluster1 | Cluster2 |
|---|---|---|
| Basal | 23.79% | 76.21% |
| Her2 | 21.79% | 78.21% |
| LumA | 69.01% | 30.99% |
| LumB | 29.59% | 70.41% |
| Normal | 86.52% | 13.48% |

**(c)** hierarchical clustering

| Subtype | Cluster1 | Cluster2 |
|---|---|---|
| Basal | 77.97% | 22.03% |
| Her2 | 76.28% | 23.72% |
| LumA | 95.53% | 4.47% |
| LumB | 82.99% | 17.01% |
| Normal | 97.75% | 2.25% |

an end node (ex. final product). Candidate subpaths are selected from the template subpaths by checking concordance between edge regulation information and correlation of two genes consisting of the edge. Then, statistical significance of a subpath is measured based the concordance edge's correlation value using the Fisher transformation. PATHOME is designed for two class (tumor vs normal) data. For the comparison with MIDAS, I performed the analysis again in two class, four cancer subtypes as one group vs. normal. In addition, there is a difference in generating pathway graphs. Since PATHOME considers only a linear path from a start node to an end node, i.e., with a constraint on graph topology, all genes are considered as separate nodes in the pathway graph. However, MIDAS considers arbitrary subpaths including those with non-linear topology, thus an original node in each KEGG pathway is a single node in the pathway graph. To handle this topological difference in two pathway graphs, the pathway graphs generated by MIDAS were used . Significance of subpath in terms of subpath length was set to 3.

**Table 4.4:** Subpath mining results of two methods. The number of subpaths determined by the two methods. The pathway that has at least one subpath is denoted as detected pathway. In 10 pathways, MIDAS predict 34 subpaths and PATHOME predict 13 subpaths. Pathway information is described here: hsa04010/MAPK signaling pathway, hsa04014/Ras signaling pathway, hsa04110/Cell cycle, hsa04151/PI3K-Akt signaling pathway, hsa04210/Apoptosis, hsa04310/Wnt signaling pathway, hsa04390/Hippo signaling pathway. hsa04550/Signaling pathways regulating pluripotency of stem cells, hsa04668/TNF signaling pathway, hsa04915/Estrogen signaling pathway

| KEGG pathway ID | Detected pathway | | # of Subpaths | |
|---|---|---|---|---|
| | MIDAS | PATHOME | MIDAS | PATHOME |
| MAPK signaling pathway | O | X | 3 | - |
| Ras signaling pathway | O | O | 5 | 4 |
| Cell cycle | O | X | 5 | - |
| PI3K-Akt signaling pathway | O | O | 4 | 5 |
| Apoptosis | O | X | 3 | - |
| Wnt signaling pathway | O | O | 4 | 3 |
| Hippo signaling pathway | O | O | 1 | 1 |
| Signaling pathways regulating pluripotency of stem cells | O | X | 4 | - |
| TNF signaling pathway | O | X | 3 | - |
| Estrogen signaling pathway | O | O | 2 | 1 |

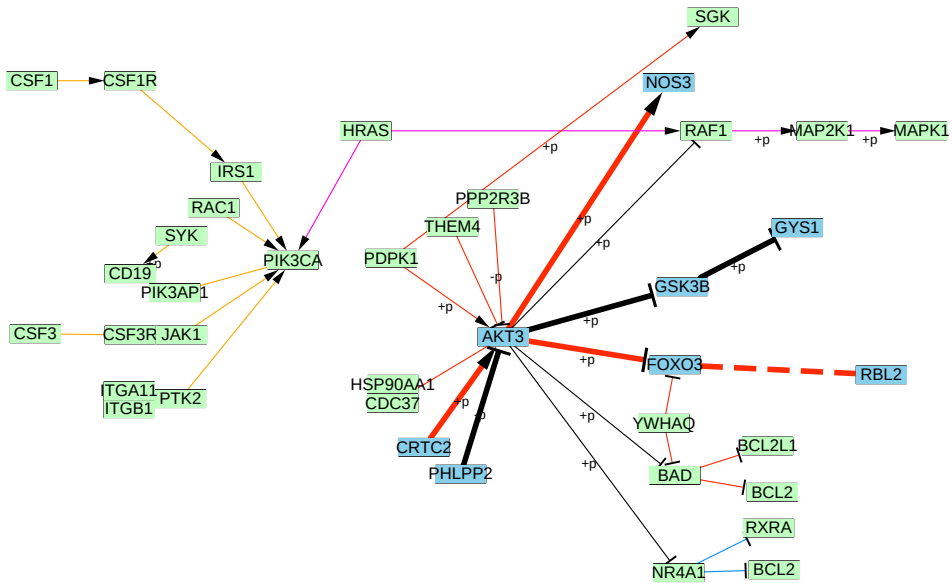



**Figure 4.4:** Comparison result on PI3K-Akt signaling pathway (hsa04151). This is merged subpaths from two methods. Subpaths extracted by PATHOME are denoted as skyblue node and concordance edges are represented as thicker edges (ex. AKT3 $-|$ GSK3$\beta$). Four subapths extracted from MIDAS also illustrated on each color: rank3 (red), rank6 (orange), rank23 (blue), rank33 (pink).

Differentially activated subpaths predicted by the two methods are summarized in Table 4.4. Among the 10 target pathways, five pathways were commonly predicted by the two methods. Among the common five pathways, subpaths of PI3K-Akt signalling pathway are illustrated in Figure 4.4. Subpaths selected by MIDAS is shown in color, and subpaths selected by PATHOME is indicated by thick edges. Although AKT3 related subpaths are commonly determined by the two methods, there is a good difference in the number of subpaths predicted and also in the number of genes in the subpaths. In terms of the number of subpaths of the common five pathways, 16 subpaths were determined by MIDAS and 14 subpaths by PATHOME. In terms of the number of genes in the subpaths, MIDAS included 310 genes while PATHOME included 83 genes. Although the number of subpaths determined by each method is no big difference, the number of genes included in the subpath were quite large. The reason why there is a big difference in the number of genes is because 14 subpaths determined by PATHOME shared many genes due to the linearity constraints of subpath topology. Also, due to the seed & expansion technique with exponential decaying threshold, MIDAS determined long length of subpaths than PATHOME: average length (MIDAS: 5.75 vs. PATHOME: 3.71) and longest length (MIDAS: 13 vs. PATHOME 5).

For all 10 pathways, MIDAS determines more subpaths than PATHOME: 34 subpaths vs. 14 subpaths. Some of subpaths detected by only MIDAS can be false positives, but they can be clues of understanding biological mechanisms. An example of pathways that were not detected by PATHOME is Cell Cycle pathway that is very well known to be important in cancer progressions. In the Cell Cycle pathway, major regulators of cell cycle progression are CDKs. For example, CDK4 interacts with many other genes, e.g. P21, RB, INK4A and 9 more genes (Keyomarsi *et al.*, 2002; Michalides *et al.*, 2002). Complex interaction mechanisms of these genes may be the reason why PATHOME

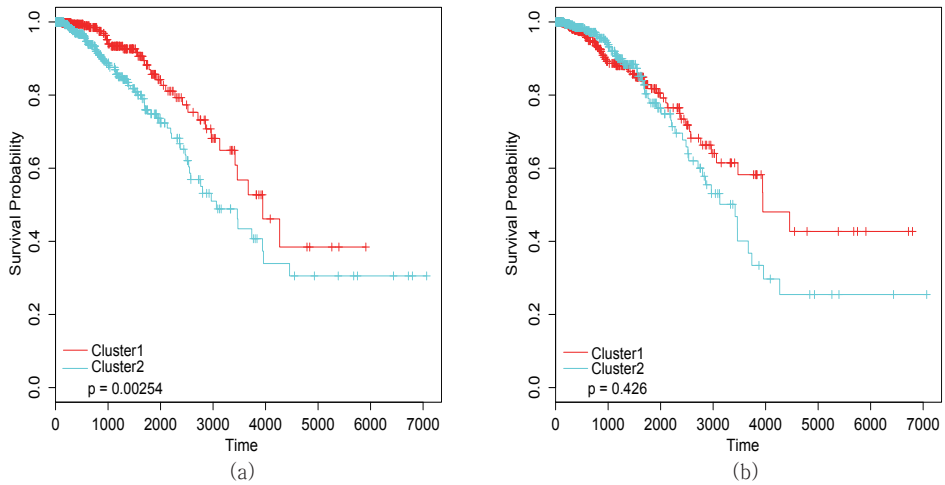



**Figure 4.5:** Survival analysis results of two methods. This is survival analysis results of two methods. (a) is survival plot of MIDAS and (b) is those of PATHOME. The statistical significance of both results are measured by log rank test from survival group generated using robust sparse K-means clustering (RSKC). Those tests are performed 10 times and median result is shown.

failed to determine cell cycle subpaths due to the linear decomposition of pathway.

To compare the prognostic power of subpaths determined by the two methods, survival analysis was performed as in Section 4.3.3. Survival tests was performed 10 times with robust sparse K-means clustering (RSKC) and the survival curves are shown in Figure 4.5. The prognosis result by MIDAS was statistically significant at the level of p-value 0.05 while the result by PATHOME was not.

### 4.3.5 Extension: detection of subnetwork on PPI network

The biological networks used in this analysis adopted from a database called pathway commons (Cerami *et al.*, 2010). Pathway Commons is a database that collects not only various PPI network databases such as BIOGRID, BIND, and REACTOME, but also information on pathway databases such as KEGG and PANTHER. The database includes regulatory networks, molecular interactions, signaling pathways, and more. The network is filtered with genes whose gene expression is measured using breast cancer transcriptome data and cancer hallmark gene set which are known to be markers of cancers (Hanahan and Weinberg, 2011), resulting in an undirected network consisting of 4,214 nodes (= genes) and 202,926 edges (= gene interactions).

Cancer subtype prediction experiments are conducted with the entire data divided by train:validation:test = 8:1:1. Through repeated experiments, the model shows reasonable classification result ranging of 78.29% to 85.39% based on weighted F1-score. This is better than MIDAS, but it is worse than the GCN+MAE model. In addition, there is a performance difference of about 7% between the lowest and highest performances, indicating that there is still a lot of work to be done. On the other hand, the GCN + MAE model uses a variety of pathways for ease of interpretation, so it takes a long time to learn several pathways individually. However, the current model has the advantage that the learning time can be significantly reduced because the features are selected on one large network. In addition, the classification performance deteriorates as abandon the fully connected layer to use CAM. This point is also observed in the original paper of CAM, and this point will be improved by referring to other studies such as grad-cam (Selvaraju *et al.*, 2017).

The figure 4.6 shows subnetwork extracted from patients belonging to each subtype. After obtaining the activation score of each node in the last CAM layer of the model, the top 10% genes are selected. The subnetworks are com-

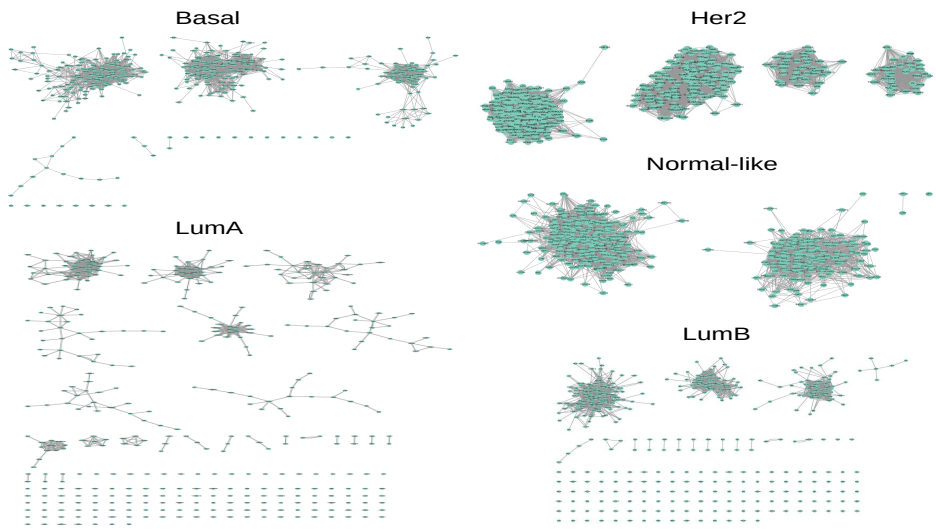



**Figure 4.6:** Subnetwork extracted from one patient belonging to each subtype by graph convolutional network and class activation map For each subtype, a subnetwork is extracted using the score of the class activation map. The extracted subnetwork is clustered through the GLay clustering algorithm considering the community structure.

posed by extracting the edges to be connected with the nodes in the original network. Then GLay, a clustering algorithm that takes into account community structures (Su *et al.*, 2010), is utilized to show interaction modules of the subnetworks. In the figure, three subtypes, Basal, Her2, and Normal-like, are grouped into one, and LumA and LumB are grouped into the other. The former group is clustered with most of the genes connected, while the latter group has a large number of genes separated from each other. This is because the former group has many interactions, that is, the genes with many edges to be selected, while the latter group seems to have selected many genes that function individually. This seems to be because the model selected the usefulness based on the node rather than the edge.

**Table 4.5:** Number of genes overlapping in each subnetwork

| | Basal | Her2 | LumA | LumB | Normal-like |
|---|---|---|---|---|---|
| Basal | - | 30 | 4 | 6 | 113 |
| Her2 | 30 | - | 0 | 59 | 98 |
| LumA | 4 | 0 | - | 26 | 57 |
| LumB | 6 | 59 | 26 | - | 0 |
| Normal-like | 113 | 98 | 57 | 0 | - |

To determine how overlapping the subnetworks of each subtype are, the number of nodes that appear in common was measured in Table4.5. LumA, which is known to have the best prognosis, showes the most overlap with Normal-like among the remaining subtypes. This is consistent with the fact that the two subtypes have the best prognosis. On the other hand, the odds are that the second most prognosis is Normal-like, and the most malignant overlaps with Basal and Her2. To address this abnormal observation, further analyses will be needed to check the hypothesis that the worst prognosis is due to abnormal expression of many normally functioning genes.

Because these subnetworks are post-processed by the activation score of the CAM layer, the entire network is actually used in the model during a classification task. Also, since the model doesnot utilize pooling layers, it does not use all the genes but only genes related to cancers. To improve this, I will devise a way of pooling considering the edge of the network, and consider the method that the subnetwork is automatically extracted as the result of the model.

# Chapter 5

# Conclusions

In my doctoral study, I proposed methods to interpret genome sequence and RNA interaction as below.

1. a new string kernel method for comparative and evolutionary comparison of DNA sequences that extends the existing $k$-spectrum string kernel by utilizing rank information and a landmark concept

2. an explainable deep learning model with graph convolutional network and attention mechanism for pathway based cancer regulation and a network propagation based bridging gaps between pathway-levels and gene-levels

3. a statistical approach and graph convolutional network method for identifying sub-modules on biological network

In the first study, I proposed the ranked $k$-spectrum string kernel for comparative and evolutionary sequence comparison. The method was based on $k$-mer frequency ranks and utilized correlations between these ranks to measure

the similarity of two sequences. The effectiveness of RKSS kernel was demonstrated through two experiments with the landmark concept. The phylogenetic tree constructed with the RKSS kernel of one landmark captured evolutionary information relatively well compared to the tree constructed with the $k$-spectrum string kernel. As the second experiment, a novel landmark space was built using the RKSS kernel with multiple landmarks. This space effectively represented the genetic properties of the three genomic regions with different characteristics. From two experiments, the relationship across information contents in exons, introns, and CpG islands was found. In terms of evolutionary information, the order of three regions was like that: exon $>$ CpG island $>$ intron. In the second study, for cancer subtype prediction using pathways, I proposed an explainable ensemble of deep pathway models. Using GCN and multi-attention, the model captured localized gene expression patterns and aggregated information spread out various pathways. On the TCGA five cancer data, the proposed method outperformed the existing pathway activity inference methods and single GCN models. In addition, unlike other methods, the proposed model used the multi-attention to obtain a list of pathways that can effectively classify and explain the characteristics of cancer subtypes from the deep learning model. Biological functions of these pathways were identified by connecting pathways and TFs by network propagation algorithm. In the final study, I designed and implemented an algorithm that determines phenotype specific subpaths and their activities. MIDAS utilized gene expression quantity information explicitly for edge activities and used a scoring scheme to measure subpath activities so that activities of multiple edges can be combined more effectively than traditional correlation-based methods. In an extensive experiment, MIDAS was successful in explaining biological mechanisms of five breast cancer subtypes. However, MIDAS did not consider the fact that a gene can belong to multiple pathways and pathways can interact with each other.

To address this issue and extend the second study, a graph convolutional network model with class activation mapping like approach on a huge biological network was designed. Although the study has not yet shown performance beyond the existing methods, it offered the possibility to extract biologically significant information beyond the limitations of the pathways. In conclusion, I developed one sequence similarity measurement for DNA sequences and two machine learning algorithms for gene expression data with biological networks. The three algorithms reduced the high dimensional features of each data to a reasonable number of features with minimizing the loss of information through biological prior knowledge.

# Bibliography

Abedin, M., Wang, D., McDonnell, M., Lehmann, U., and Kelekar, A. (2007). Autophagy delays apoptotic death in breast cancer cells following dna damage. *Cell Death & Differentiation*, **14**(3), 500–510.

Alcaraz, N., List, M., Batra, R., Vandin, F., Ditzel, H. J., and Baumbach, J. (2017). De novo pathway-based biomarker identification. *Nucleic acids research*, **45**(16), e151–e151.

Allman, E. S., Rhodes, J. A., and Sullivant, S. (2017). Statistically consistent k-mer methods for phylogenetic tree reconstruction. *Journal of Computational Biology*, **24**(2), 153–171.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

Atwood, J. and Towsley, D. (2016). Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1993–2001.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, **12**(1), 56–68.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS computational biology*, **4**(10), e1000173.

Berezikov, E., Guryev, V., Plasterk, R. H., and Cuppen, E. (2004). Conreal: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome research*, **14**(1), 170–178.

Berns, K., Horlings, H. M., Hennessy, B. T., Madiredjo, M., Hijmans, E. M., Beelen, K., Linn, S. C., Gonzalez-Angulo, A. M., Stemke-Hale, K., Hauptmann, M., *et al.* (2007). A functional genetic approach identifies the pi3k pathway as a major determinant of trastuzumab resistance in breast cancer. *Cancer cell*, **12**(4), 395–402.

Bharadwaj, R. and Yu, H. (2004). The spindle checkpoint, aneuploidy, and cancer. *Oncogene*, **23**(11), 2016–2027.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**(8), 1091–1093.

Biswas, D. K., Cruz, A. P., Gansberger, E., and Pardee, A. B. (2000). Epidermal growth factor-induced nuclear factor $\kappa$b activation: a major pathway of cell-cycle progression in estrogen-receptor negative breast cancer cells. *Proceedings of the National Academy of Sciences*, **97**(15), 8542–8547.

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences*, **83**(14), 5155–5159.

Bonham-Carter, O., Steele, J., and Bastola, D. (2013). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics*, **15**(6), 890–905.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. (2012). Rfam 11.0: 10 years of rna families. *Nucleic acids research*, **41**(D1), D226–D232.

Cakir, A., Gonul, I. I., and Uluoglu, O. (2012). A comprehensive morphological study for basal-like breast carcinomas with comparison to nonbasal-like carcinomas. *Diagnostic pathology*, **7**(1), 145.

Casimiro, M. C., Crosariol, M., Loro, E., Li, Z., and Pestell, R. G. (2012). Cyclins and cell cycle control in cancer and disease. *Genes & cancer*, **3**(11-12), 649–657.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, **538**(7623), 20.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2010). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, **39**(suppl_1), D685–D690.

Chae, H., Park, J., Lee, S.-W., Nephew, K. P., and Kim, S. (2013). Comparative analysis using k-mer and k-flank patterns provides evidence for cpg island sequence evolution in mammalian genomes. *Nucleic acids research*, **41**(9), 4783–4791.

Chang, J. T., Carvalho, C., Mori, S., Bild, A. H., Gatza, M. L., Wang, Q., Lucas, J. E., Potti, A., Febbo, P. G., West, M., *et al.* (2009). A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Molecular cell*, **34**(1), 104–114.

Chen, D., Sun, Y., Wei, Y., Zhang, P., Rezaeian, A. H., Teruya-Feldstein, J., Gupta, S., Liang, H., Lin, H.-K., Hung, M.-C., *et al.* (2012). Lifr is a breast cancer metastasis suppressor upstream of the hippo-yap pathway and a prognostic marker. *Nature medicine*, **18**(10), 1511–1517.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551.

Croft, D., O'kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., *et al.* (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, **39**(suppl_1), D691–D697.

Cuturi, M. and Vert, J.-P. (2005). The context-tree kernel for strings. *Neural Networks*, **18**(8), 1111–1123.

Czabotar, P. E., Lessene, G., Strasser, A., and Adams, J. M. (2014). Control of apoptosis by the bcl-2 protein family: implications for physiology and therapy. *Nature reviews Molecular cell biology*, **15**(1), 49–63.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast

cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, **5**(10), 2929.

Das, S., Deb, T., Dey, N., Ashour, A. S., Bhattacharya, D., and Tibarewala, D. (2018). Optimal choice of k-mer in composition vector method for genome sequence comparison. *Genomics*, **110**(5), 263–273.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852.

Dhillon, I. S., Guan, Y., and Kulis, B. (2007). Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, **29**(11).

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, **110**(16), 6388–6393.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Eichbaum, C., Meyer, A.-S., Wang, N., Bischofs, E., Steinborn, A., Bruckner, T., Brodt, P., Sohn, C., and Eichbaum, M. H. (2011). Breast cancer cell-derived cytokines, macrophages and cell adhesion: implications for metastasis. *Anticancer research*, **31**(10), 3219–3227.

Esquivel-Velázquez, M., Ostoa-Saloma, P., Palacios-Arreola, M. I., Nava-Castro, K. E., Castro, J. I., and Morales-Montor, J. (2015). The role of cytokines in breast cancer development and progression. *Journal of Interferon & Cytokine Research*, **35**(1), 1–16.

Fabre, J., Giustinniani, J., Garbar, C., Merrouche, Y., Antonicelli, F., and Bensussan, A. (2018). The interleukin-17 family of cytokines in breast cancer. *International journal of molecular sciences*, **19**(12), 3880.

Fong, Y., Datta, S., Georgiev, I. S., Kwong, P. D., and Tomaras, G. D. (2014). Kernel-based logistic regression model for protein sequence without vectorialization. *Biostatistics*, **16**(3), 480–492.

Forêt, S., Wilson, S. R., and Burden, C. J. (2009). Characterizing the d2 statistic: word matches in biological sequences. *Statistical applications in genetics and molecular biology*, **8**(1), 1–21.

García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Frontiers in physiology*, **6**.

Gatza, M. L., Lucas, J. E., Barry, W. T., Kim, J. W., Wang, Q., Crawford, M. D., Datto, M. B., Kelley, M., Mathey-Prevot, B., Potti, A., *et al.* (2010). A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences*, **107**(15), 6994–6999.

Gish, W. and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature genetics*, **3**(3), 266.

Glaab, E., Garibaldi, J. M., and Krasnogor, N. (2010). Learning pathway-based decision rules to classify microarray cancer samples. In *German Conference on Bioinformatics 2010*. Gesellschaft für Informatik eV.

Grimm, L. J., Johnson, K. S., Marcom, P. K., Baker, J. A., and Soo, M. S. (2014). Can breast cancer molecular subtype help to select patients for preoperative mr imaging? *Radiology*, **274**(2), 352–358.

Gunning, D. (2017). Explainable artificial intelligence (xai).

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, **30**(2), 129–150.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., *et al.* (2017). Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, **46**(D1), D380–D386.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, **144**(5), 646–674.

Hartman, Z. C., Poage, G. M., Den Hollander, P., Tsimelzon, A., Hill, J., Panupinthu, N., Zhang, Y., Mazumdar, A., Hilsenbeck, S. G., Mills, G. B., *et al.* (2013). Growth of triple-negative breast cancer cells relies upon coordinate autocrine expression of the proinflammatory cytokines il-6 and il-8. *Cancer research*, **73**(11), 3470–3480.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report.

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015). Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC bioinformatics*, **16**(1), 169.

Hennessy, B. T., Gonzalez-Angulo, A.-M., Stemke-Hale, K., Gilcrease, M. Z., Krishnamurthy, S., Lee, J.-S., Fridlyand, J., Sahin, A., Agarwal, R., Joy, C., *et al.* (2009). Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer research*, **69**(10), 4116–4124.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature methods*, **10**(11), 1108.

Howe, L. R. and Brown, A. M. (2004). Wnt signaling and breast cancer. *Cancer biology & therapy*, **3**(1), 36–41.

Huising, M. O., Geven, E. J., Kruiswijk, C. P., Nabuurs, S. B., Stolte, E. H., Spanings, F. T., Verburg-van Kemenade, B. L., and Flik, G. (2006). Increased leptin expression in common carp (cyprinus carpio) after food intake but not after fasting or feeding to satiation. *Endocrinology*, **147**(12), 5786–5797.

Hwang, K.-T., Kim, J., Jung, J., Chang, J. H., Chai, Y. J., Oh, S. W., Oh, S., Kim, Y. A., Park, S. B., and Hwang, K. R. (2019). Impact of breast cancer subtypes on prognosis of women with operable invasive breast cancer: a population-based study using seer database. *Clinical Cancer Research*, **25**(6), 1970–1979.

Jin, L., Zuo, X.-Y., Su, W.-Y., Zhao, X.-L., Yuan, M.-Q., Han, L.-Z., Zhao, X., Chen, Y.-D., and Rao, S.-Q. (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*, **12**(5), 210–220.

Jo, K., Jung, I., Moon, J. H., and Kim, S. (2016). Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways. *Bioinformatics*, **32**(12), i128–i136.

Judeh, T., Johnson, C., Kumar, A., and Zhu, D. (2013). Teak: topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic acids research*, **41**(3), 1425–1437.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Katoh, M. and Katoh, M. (2007). Wnt signaling pathway and stem cell signaling network. *Clinical Cancer Research*, **13**(14), 4042–4045.

Kelder, T., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2010). Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS Biol*, **8**(8), e1000472.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, **12**(6), 996–1006.

Keyomarsi, K., Tucker, S. L., Buchholz, T. A., Callister, M., Ding, Y., Hortobagyi, G. N., Bedrosian, I., Knickerbocker, C., Toyofuku, W., Lowe, M., *et al.* (2002). Cyclin e and survival in patients with breast cancer. *New England Journal of Medicine*, **347**(20), 1566–1575.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, **8**(2), e1002375.

Kim, G., Ouzounova, M., Quraishi, A. A., Davis, A., Tawakkol, N., Clouthier, S. G., Malik, F., Paulson, A. K., D'Angelo, R. C., Korkaya, S., *et al.* (2015). Socs3-mediated regulation of inflammatory cytokines in pten and p53 inactivated triple negative breast cancer model. *Oncogene*, **34**(6), 671.

Kim, S., Kon, M., and DeLisi, C. (2012). Pathway-based classification of cancer subtypes. *Biology direct*, **7**(1), 21.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR 2017*.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, **82**(4), 949–958.

Kondo, Y., Salibian-Barrera, M., and Zamar, R. (2016). Rskc: An r package for a robust and sparse k-means clustering algorithm. *Journal of Statistical Software*, **72**(5), 1–26.

Kong, Y. and Yu, T. (2018). A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics*, **34**(21), 3727–3737.

Koumakis, L., Kanterakis, A., Kartsaki, E., Chatzimina, M., Zervakis, M., Tsiknakis, M., Vassou, D., Kafetzopoulos, D., Marias, K., Moustakis, V., *et al.* (2016). Minepath: Mining for phenotype differential sub-paths in molecular pathways. *PLoS Comput Biol*, **12**(11), e1005187.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, **44**(W1), W90–W97.

Kumar, S. (2007). Caspase function in programmed cell death. *Cell Death & Differentiation*, **14**(1), 32–43.

Kunz, M., Jeromin, J., Fuchs, M., Christoph, J., Veronesi, G., Flentje, M., Nietzer, S., Dandekar, G., and Dandekar, T. (2019). In silico signaling modeling to understand cancer pathways and treatment responses. *Briefings in bioinformatics*.

Lai, D., Ho, K. C., Hao, Y., and Yang, X. (2011). Taxol resistance in breast cancer cells is mediated by the hippo pathway component taz and its down-

stream transcriptional targets cyr61 and ctgf. *Cancer research*, **71**(7), 2728–2738.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, **172**(4), 650–665.

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., *et al.* (2007). Clustal w and clustal x version 2.0. *bioinformatics*, **23**(21), 2947–2948.

Lee, S., Park, Y., and Kim, S. (2017). Midas: Mining differentially activated subpaths of kegg pathways from multi-class rna-seq data. *Methods*, **124**, 13–24.

Lee, S., Lim, S., Lee, T., Noh, Sung, I., and Kim, S. (2019a). Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics (Under Revision)*.

Lee, S., Lee, T., Noh, Y.-K., and Kim, S. (2019b). Ranked k-spectrum kernel for comparative and evolutionary comparison of exons, introns, and cpg islands. *IEEE/ACM transactions on computational biology and bioinformatics*.

Leiserson, M. D., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., *et al.* (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, **47**(2), 106.

Leslie, C., Eskin, E., and Noble, W. S. (2001). The spectrum kernel: A string

kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**(4), 467–476.

Li, J., Zhang, J., Chen, H., Chen, X., Lan, J., and Liu, C. (2013). Complete mitochondrial genome of the medicinal mushroom ganoderma lucidum. *PLoS One*, **8**(8), e72038.

Li, Y., Welm, B., Podsypanina, K., Huang, S., Chamorro, M., Zhang, X., Rowlands, T., Egeblad, M., Cowin, P., Werb, Z., *et al.* (2003). Evidence that transgenes encoding components of the wnt signaling pathway preferentially induce mammary cancers from progenitor cells. *Proceedings of the National Academy of Sciences*, **100**(26), 15853–15858.

Lim, S., Park, Y., Hur, B., Kim, M., Han, W., and Kim, S. (2016). Protein interaction network (pin)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods*.

Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*.

Liu, Z., Li, M., Jiang, Z., and Wang, X. (2018). A comprehensive immunologic portrait of triple-negative breast cancer. *Translational oncology*, **11**(2), 311–329.

Lo, T. L., Yusoff, P., Fong, C. W., Guo, K., McCaw, B. J., Phillips, W. A., Yang, H., Wong, E. S. M., Leong, H. F., Zeng, Q., *et al.* (2004). The ras/mitogen-activated protein kinase pathway inhibitor and likely tumor

suppressor proteins, sprouty 1 and sprouty 2 are deregulated in breast cancer. *Cancer research*, **64**(17), 6127–6136.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, **2**(Feb), 419–444.

Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I., and Xiong, M. (2010). Genome-wide gene and pathway analysis. *European Journal of Human Genetics*, **18**(9), 1045–1053.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, **10**(1), 161.

Mao, L., Yuan, L., Slakey, L. M., Jones, F. E., Burow, M. E., and Hill, S. M. (2010). Inhibition of breast cancer cell invasion by melatonin is mediated through regulation of the p38 mitogen-activated protein kinase signaling pathway. *Breast Cancer Research*, **12**(6), R107.

Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. (2012). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic acids research*, **41**(1), e19–e19.

Massarweh, S., Osborne, C. K., Creighton, C. J., Qin, L., Tsimelzon, A., Huang, S., Weiss, H., Rimawi, M., and Schiff, R. (2008). Tamoxifen resistance in breast tumors is driven by growth factor receptor signaling with repression of classic estrogen receptor genomic function. *Cancer research*, **68**(3), 826–833.

Menendez, J. A., Vellon, L., Mehmi, I., Teng, P. K., Griggs, D. W., and Lupu, R. (2005). A novel cyr61-triggered 'cyr61-$\alpha$v$\beta$3 integrin loop'regulates

breast cancer cell survival and chemosensitivity through activation of erk1/erk2 mapk signaling pathway. *Oncogene*, **24**(5), 761–779.

Mi, H., Muruganujan, A., and Thomas, P. D. (2013). Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, **41**(D1), D377–D386.

Michalides, R., Van Tinteren, H., Balkenende, A., Vermorken, J., Benraadt, J., Huldij, J., and Van Diest, P. (2002). Cyclin a is a prognostic indicator in early stage breast cancer with and without tamoxifen treatment. *British journal of cancer*, **86**(3), 402–408.

Middleton, S. A. and Kim, J. (2014). Nofold: Rna structure clustering without folding or alignment. *RNA*.

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., *et al.* (2007). 28-way vertebrate alignment and conservation track in the ucsc genome browser. *Genome research*, **17**(12), 000–000.

Mirzoeva, O. K., Das, D., Heiser, L. M., Bhattacharya, S., Siwak, D., Gendelman, R., Bayani, N., Wang, N. J., Neve, R. M., Guan, Y., *et al.* (2009). Basal subtype and mapk/erk kinase (mek)-phosphoinositide 3-kinase feedback signaling determine susceptibility of breast cancer cells to mek inhibition. *Cancer research*, **69**(2), 565–572.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124.

Moon, J. H., Lim, S., Jo, K., Lee, S., Seo, S., and Kim, S. (2017). Pintnet:

construction of condition-specific pathway interaction network by computing shortest paths on weighted ppi. *BMC systems biology*, **11**(2), 15.

Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., and Warthmann, N. (2017). kwip: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS computational biology*, **13**(9), e1005727.

Nam, S., Chang, H. R., Kim, K.-T., Kook, M.-C., Hong, D., Kwon, C., Jung, H. R., Park, H. S., Powis, G., Liang, H., *et al.* (2014). Pathome: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene*, **33**(41), 4941.

Nersisyan, L., Samsonyan, R., and Arakelyan, A. (2014). Cykeggparser: tailoring kegg pathways to fit into systems biology analysis workflows. *F1000Research*, **3**.

Network, C. G. A. *et al.* (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61.

Nojoomi, S. and Koehl, P. (2017a). String kernels for protein sequence comparisons: improved fold recognition. *BMC bioinformatics*, **18**(1), 137.

Nojoomi, S. and Koehl, P. (2017b). A weighted string kernel for protein fold recognition. *BMC bioinformatics*, **18**(1), 378.

Osborne, C. K., Shou, J., Massarweh, S., and Schiff, R. (2005). Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer. *Clinical cancer research*, **11**(2), 865s–870s.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., *et al.* (2005). The subsystems approach to genome annotation and its use in the

project to annotate 1000 genomes. *Nucleic acids research*, **33**(17), 5691–5702.

Papaliagkas, V., Anogianaki, A., Anogianakis, G., and Ilonidis, G. (2007). The proteins and the mechanisms of apoptosis: A mini-review of the fundamentals. *Hippokratia*, **11**(3), 108–113.

Paplomata, E. and O'Regan, R. (2014). The pi3k/akt/mtor pathway in breast cancer: targets, trials and biomarkers. *Therapeutic advances in medical oncology*, **6**(4), 154–166.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, **27**(8), 1160.

Pearson, K. (1905). The problem of the random walk. *Nature*, **72**(1867), 342.

Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., Michelini, K., Zehr, S., Yoder, A. D., Stephens, M., *et al.* (2012). Comparative rna sequencing reveals substantial genetic variation in endangered primates. *Genome research*, **22**(4), 602–610.

Porter, P. L., Malone, K. E., Heagerty, P. J., Alexander, G. M., Gatti, L. A., Firpo, E. J., Daling, J. R., and Roberts, J. M. (1997). Expression of cell-cycle regulators p27kip1 and cyclin e, alone and in combination, correlate with survival in young breast cancer patients. *Nature medicine*, **3**(2), 222–225.

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., *et al.* (2015). Ndex, the network data exchange. *Cell systems*, **1**(4), 302–305.

Rätsch, G., Sonnenburg, S., and Schölkopf, B. (2005). Rase: recognition of alternatively spliced exons in c. elegans. *Bioinformatics*, **21**(suppl_1), i369–i377.

Reis-Filho, J. S. and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, **378**(9805), 1812–1823.

Rhee, S., Seo, S., and Kim, S. (2018). Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3527–3534. AAAI Press.

Robinson, O., Dylus, D., and Dessimoz, C. (2016). Phylo. io: interactive viewing and comparison of large phylogenetic trees on the web. *Molecular biology and evolution*, **33**(8), 2163–2166.

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, **20**(11), 1682–1689.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4**(4), 406–425.

Santen, R. J., Song, R. X., McPherson, R., Kumar, R., Adam, L., Jeng, M.-H., and Yue, W. (2002). The role of mitogen-activated protein (map) kinase in breast cancer. *The Journal of steroid biochemistry and molecular biology*, **80**(2), 239–256.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., *et al.* (2005). An integra-

tive genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, **37**(7), 710.

Schlange, T., Matsuda, Y., Lienhard, S., Huber, A., and Hynes, N. E. (2007). Autocrine wnt signaling contributes to breast cancer cell proliferation via the canonical wnt pathway and egfr transactivation. *Breast cancer research*, **9**(5), R63.

Seeger, M. (2002). Covariance kernels from bayesian generative models. In *Advances in neural information processing systems*, pages 905–912.

Segura-Lepe, M. P., Keun, H. C., and Ebbels, T. M. (2019). Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC bioinformatics*, **20**(1), 543.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Sequencing, T. M. G., Worley, K. C., Warren, W. C., Rogers, J., Locke, D., Muzny, D. M., Mardis, E. R., Weinstock, G. M., Tardif, S. D., Aagaard, K. M., *et al.* (2014). The common marmoset genome provides insight into primate biology and evolution. *Nature genetics*, **46**(8), 850.

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

Shen, W.-J., Wong, H.-S., Xiao, Q.-W., Guo, X., and Smale, S. (2014). In-

troduction to the peptide binding problem of computational immunology: New results. *Foundations of Computational Mathematics*, **14**(5), 951–984.

Smith, T. and BEYER, W. (1976). M. s. waterman.

Smola, A. J. and Vishwanathan, S. (2003). Fast kernels for string and tree matching. In *Advances in neural information processing systems*, pages 585–592.

Söding, J. (2004). Protein homology detection by hmm–hmm comparison. *Bioinformatics*, **21**(7), 951–960.

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., and Sun, F. (2013). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, **15**(3), 343–353.

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, **100**(18), 10393–10398.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, **34**(suppl_1), D535–D539.

Steinegger, M. and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature communications*, **9**(1), 2542.

Stewart, B., Wild, C. P., *et al.* (2016). World cancer report 2014. *World*.

Stuart, G. W., Moffett, K., and Baker, S. (2002). Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **18**(1), 100–108.

Stuelten, C. H., Byfield, S. D., Arany, P. R., Karpova, T. S., Stetler-Stevenson, W. G., and Roberts, A. B. (2005). Breast cancer cells induce stromal fibroblasts to express mmp-9 via secretion of tnf-$\alpha$ and tgf-$\beta$. *Journal of cell science*, **118**(10), 2143–2153.

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., and Meng, F. (2010). Glay: community structure analysis of biological networks. *Bioinformatics*, **26**(24), 3135–3137.

Su, J., Yoon, B.-J., and Dougherty, E. R. (2009). Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PloS one*, **4**(12), e8161.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 75–82.

Thomas, P., Pang, Y., Filardo, E., and Dong, J. (2005). Identity of an estrogen membrane receptor coupled to a g protein in human breast cancer cells. *Endocrinology*, **146**(2), 624–632.

Tokunaga, E., Kimura, Y., Mashino, K., Oki, E., Kataoka, A., Ohno, S.,

Morita, M., Kakeji, Y., Baba, H., and Maehara, Y. (2006). Activation of pi3k/akt signaling and hormone resistance in breast cancer. *Breast cancer*, **13**(2), 137–144.

Ulitsky, I., Burstein, D., Tuller, T., and Chor, B. (2006). The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, **13**(2), 336–350.

van de Sande, W. W. (2012). Phylogenetic analysis of the complete mitochondrial genome of madurella mycetomatis confirms its taxonomic position within the order sordariales. *PLoS One*, **7**(6), e38654.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, **26**(12), i237–i245.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vinga, S. (2007). Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for dna and protein classification. *Advanced Computational Methods for Biocomputing and Bioimaging*, pages 71–107.

Vinga, S. (2014). Alignment-free methods in computational biology.

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics*, **19**(4), 513–523.

Viswanathan, G. A., Seto, J., Patil, S., Nudelman, G., and Sealfon, S. C.

(2008). Getting started in biological pathway construction and analysis. *PLoS Computational Biology*, **4**(2), e16.

Wang, Y., Yu, Y., Tsuyada, A., Ren, X., Wu, X., Stubblefield, K., Rankin-Gee, E. K., and Wang, S. E. (2011). Transforming growth factor-$\beta$ regulates the sphere-initiating stem cell-like feature in breast cancer through mirna-181 and atm. *Oncogene*, **30**(12), 1470–1480.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, **10**(1), 57–63.

Watkins, C. (1999). Dynamic alignment kernels. In *Advances in neural information processing systems*, pages 39–50.

Wu, T.-J., Hsieh, Y.-C., and Li, L.-A. (2001). Statistical measures of dna sequence dissimilarity under markov chain models of base composition. *Biometrics*, **57**(2), 441–448.

Yang, J., Liao, D., Chen, C., Liu, Y., Chuang, T.-H., Xiang, R., Markowitz, D., Reisfeld, R. A., and Luo, Y. (2013). Tumor-associated macrophages regulate murine breast cancer stem cells through a novel paracrine egfr/stat3/sox-2 signaling pathway. *Stem cells*, **31**(2), 248–258.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., *et al.* (2003). Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, **4**(4), R28.

Zhang, J. D. and Wiemann, S. (2009). Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, **25**(11), 1470–1471.

Zhang, W., Ma, J., and Ideker, T. (2018). Classifying tumors by supervised network propagation. *Bioinformatics*, **34**(13), i484–i493.

Zheng, W., Lin, H., Luo, L., Zhao, Z., Li, Z., Zhang, Y., Yang, Z., and Wang, J. (2017). An attention-based effective neural model for drug-drug interactions extraction. *BMC bioinformatics*, **18**(1), 445.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *arXiv preprint arXiv:1802.00543*.

Zubaer, A., Wai, A., and Hausner, G. (2018). The mitochondrial genome of endoconidiophora resinifera is intron rich. *Scientific reports*, **8**(1), 17591.

# 국문초록

생물체 간 표현형의 차이는 각 개체의 유전적 정보 차이로부터 기인한다. 유전적 정보의 변화에 따라서, 각 생물체는 서로 다른 종으로 진화하기도 하고, 같은 병에 걸린 환자라도 서로 다른 예후를 보이기도 한다. 이처럼 중요한 생물학적 정보는 대용량 시퀀싱 분석 기법 등을 통해 다양한 오믹스 데이터로 측정된다. 그러나, 오믹스 데이터는 고차원 특징 및 소규모 표본 데이터이기 때문에, 오믹스 데이터로부터 생물학적 정보를 해석하는 것은 매우 어려운 문제이다. 일반적으로, 데이터 특징의 개수가 샘플의 개수보다 많을 때, 오믹스 데이터의 해석을 가장 난해한 기계학습 문제들 중 하나로 만듭니다.

본 박사학위 논문은 기계학습 기법을 활용하여 고차원적인 생물학적 데이터로부터 생물학적 정보를 추출하기 위한 새로운 생물정보학 방법들을 고안하는 것을 목표로 한다.

첫 번째 연구는 DNA 서열을 활용하여 종 간 비교와 동시에 DNA 서열상에 있는 다양한 지역에 담긴 생물학적 정보를 유전적 관점에서 해석해보고자 하였다. 이를 위해, 순위 기반 $k$ 단어 문자열 비교방법, RKSS 커널을 개발하여 다양한 게놈 상의 지역에서 여러 종 간 비교 실험을 수행하였다. RKSS 커널은 기존의 $k$ 단어 문자열 커널을 확장한 것으로, $k$ 길이 단어의 순위 정보와 종 간 공통점을 표현하는 비교기준점 개념을 활용하였다. $k$ 단어 문자열 커널은 $k$의 길이에 따라 단어 수가 급증하지만, 비교기준점은 극소수의 단어로 이루어져 있으므로 서열 간 유사도를 계산하는 데 필요한 계산량을 효율적으로 줄일 수 있다. 게놈 상의 세 지역에 대해서 실험을 진행한 결과, RKSS 커널은 기존의 커널에 비해 종 간 유사도 및 차이를 효율적으로 계산할 수 있었다. 또한, RKSS 커널은 실험에 사용된 생물학적 지역에 포함된 생물학적 정보량 차이를 생물학적 지식과 부합되는 순서로 비교할 수 있었다.

두 번째 연구는 생물학적 네트워크를 통해 복잡하게 얽힌 유전자 상호작용 간 정보를 해석하여, 더 나아가 생물학적 기능 해석을 통해 암의 아형을 분류하고자 하였다. 이를 위해, 그래프 컨볼루션 네트워크와 어텐션 메커니즘을 활용하여 패스웨이 기반 해석 가능한 암 아형 분류 모델(GCN+MAE)을 고안하였다. 그래프 컨볼루션 네트워크를 통해서 생물학적 사전 지식인 패스웨이 정보를 학습하여 복잡한 유전자 상호작용 정보를 효율적으로 다루었다. 또한, 여러 패스웨이 정보를 어텐션 메커니즘을 통해 해석 가능한 수준으로 병합하였다. 마지막으로, 학습한 패스웨이 레벨 정보를 보다 복잡하고 다양한 유전자 레벨로 효율적으로 전달하기 위해서 네트워크 전파 알고리즘을 활용하였다. 다섯 개의 암 데이터에 대해 GCN+MAE 모델을 적용한 결과, 기존의 암 아형 분류 모델들보다 나은 성능을 보였으며 암 아형 특이적인 패스웨이 및 생물학적 기능을 발굴할 수 있었다.

세 번째 연구는 패스웨이로부터 서브 패스웨이/네트워크를 찾기 위한 연구다. 패스웨이나 생물학적 네트워크에 단일 생물학적 기능이 아니라 다양한 생물학적 기능이 포함되어 있음에 주목하였다. 단일 기능을 지닌 유전자 조합을 찾기 위해서 생물학적 네트워크상에서 조건 특이적인 유전자 모듈을 찾고자 하였으며 MIDAS라는 도구를 개발하였다. 패스웨이로부터 유전자 상호작용 간 활성도를 유전자 발현량과 네트워크 구조를 통해 계산하였다. 계산된 활성도들을 활용하여 다중 클래스에서 서로 다르게 활성화된 서브 패스들을 통계적 기법에 기반하여 발굴하였다. 또한, 어텐션 메커니즘과 그래프 컨볼루션 네트워크를 통해서 해당 연구를 패스웨이보다 더 큰 생물학적 네트워크에 확장하려고 시도하였다. 유방암 데이터에 대해 실험을 진행한 결과, MIDAS와 딥러닝 모델을 다중 클래스에서 차이가 나는 유전자 모듈을 효과적으로 추출할 수 있었다.

결론적으로, 본 박사학위 논문은 DNA 서열에 담긴 진화적 정보량 비교, 패스웨이 기반 암 아형 분류, 조건 특이적인 유전자 모듈 발굴을 위한 새로운 기계학습 기법을 제안하였다.

**주요어**: 고차원 데이터, 생물학적 사전지식, DNA 서열, 유전자 발현량, 기계학습
**학번**: 2014-21754