理學博士學位論文

Hypergraph Models for Identifying Co-Regulatory Genomic Interactions

동시조절 유전적 상호작용 발굴을 위한 하이퍼그래프 모델

2014年 2月

서울大學校 大學院

협동과정 생물정보학

金 秀 珍

Hypergraph Models for Identifying Co-Regulatory Genomic Interactions

동시조절 유전적 상호작용 발굴을 위한 하이퍼그래프 모델

指導教授　張 炳 卓

이 論文을 理學博士 學位論文으로 提出함

2013年 10月

서울大學校 大學院

협동과정 생물정보학

金 秀 珍

金秀珍의 理學博士 學位論文을 認准함

2013年 11月

委員長 李 柄 宇

副委員長 張 炳 卓

委　員 尹 晟 老

委　員 鄭 東 均

委　員 申 守 容

Hypergraph Models for Identifying Co-Regulatory Genomic Interactions

동시조절 유전적 상호작용 발굴을 위한 하이퍼그래프 모델

Soo-Jin Kim

Ph.D. Thesis

Interdisciplinary Program in Bioinformatics
Seoul National University

February 2014

Supervisor: Byoung-Tak Zhang

Abstract

A comprehensive understanding of biological systems requires the analysis of higher-order interactions among many genomic factors. Various genomic factors cooperate to affect biological processes including cancer occurrence, progression and metastasis. However, the complexity of genomic interactions presents a major barrier to identifying their co-regulatory roles and functional effects. Thus, this dissertation addresses the problem of analyzing complex relationships among many genomic factors in biological processes including cancers. We propose a hypergraph approach for modeling, learning and extracting: explicitly modeling higher-order genomic interactions, efficiently learning based on evolutionary methods, and effectively extracting biological knowledge from the model.

A hypergraph model is a higher-order graphical model explicitly representing complex relationships among many variables from high-dimensional data. This property allows the proposed model to be suitable for the analysis of biological and medical phenomena characterizing higher-order interactions between various genomic factors. This dissertation proposes the advanced hypergraph-based models in terms of the learning methods and the model structures to analyze large-scale biological data focusing on identifying co-regulatory genomic interactions on a genome-wide level. We introduce an evolutionary approach based on information-theoretic criteria into the learning mechanisms for efficiently searching a huge problem space reflecting higher-order interactions between factors. This evolutionary learning is explained from the perspective of a sequential Bayesian sampling framework. Also, a hierarchy is introduced into the hypergraph model for modeling hierarchical genomic relationships. This hierarchical structure allows the hypergraph model to explicitly represent gene regulatory circuits as functional blocks or groups across the level of epigenetic, transcriptional, and post-transcriptional regulation. Moreover, the proposed graph-analyzing method is able to grasp the global structures of biological systems such as genomic modules and regulatory networks by analyzing the learned model structures.

The proposed model is applied to analyzing cancer genomics considered as a major topic in current biology and medicine. We show that the performance of our model competes with or outperforms state-of-the-art models on multiple cancer genomic data. Furthermore, the propose model is capable of discovering new or hidden patterns as candidates of potential gene regulatory circuits such as gene modules, miRNA-mRNA networks, and multiple genomic interactions, associated with the specific cancer. The results of these analysis can provide several crucial evidences that can pave the way for identifying unknown functions in the cancer system. The proposed hypergraph model will contribute to elucidating core regulatory mechanisms and to comprehensive understanding of biological processes including cancers.

Keywords: **Hypergraph, Higher-order graphical model,**
Evolutionary learning, Genomic interaction,
Gene module, miRNA-mRNA network, Cancer

Student Number: 2005-20561

Contents

List of Tables

List of Figures

Chapter 1

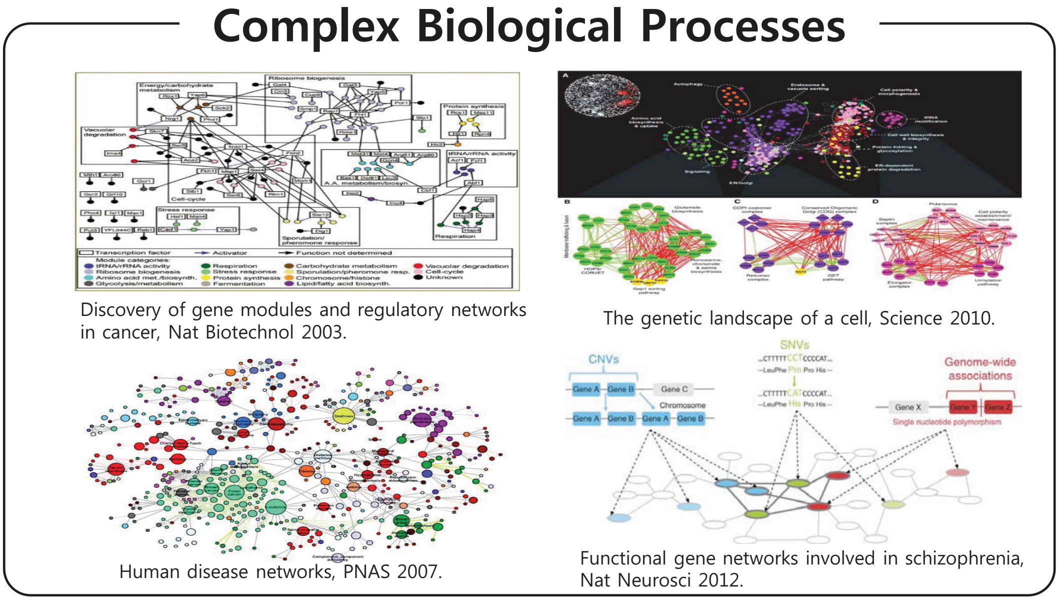
Introduction

1.1 Background and Motivation

Recent biological and medical research advances from studying simple to complex traits, from carrying out separated to integrated analyzes of different genomic data sources, from analyzing a single gene to multiple gene interactions at the systems level. Computational approaches, which analyze gene regulatory relationships on a genome-wide scale from high-throughput data, have led to a deluge of systemic insights into a variety of biological and medical areas.

Cancer is one of the important challenges in biology and medicine because it is still the lethal disease of the leading cause of death worldwide. High-throughput data have been massively produced to understand cancer mechanisms for more several years. Despite such efforts, the mechanism of cancer is not clearly deciphered yet.

The regulation of cancer is a complicated phenomenon, induced by complex interactions among various genetic factors. It is mostly related to modular construction and combinatorial control by multiple genetic factors such as miRNAs



Complex Biological Processes

Discovery of gene modules and regulatory networks in cancer, Nat Biotechnol 2003.

The genetic landscape of a cell, Science 2010.

Human disease networks, PNAS 2007.

Functional gene networks involved in schizophrenia, Nat Neurosci 2012.

Figure 1.1: Analyzing higher-order genomic interactions is necessary to understand complex and various biological processes including cancer.

and mRNAs across the transcriptional, post-transcriptional and epigenetic levels. Thus, elucidating multiple genomic interactions at multicellular level is essential to understand complex biological processes including cancer development and progression more precisely (Figure 1.1). Furthermore, it can provide new insights into the behavior of complex biological systems. However, the analysis of higher-order relationships between many genetic factors is rendered as a challenging problem due to the complexity of their interactions.

Herein, one of the major issues associated with investigating complex genomic interactions is the volume of data to be analyzed; as the number of genes increases the number of potential interactions increases exponentially, known as ’curse of dimensionality' (Moore and Ritchie, 2004). The potential complexity of such interactions supports the use of various machine learning techniques for analyzing co-regulatory genomic relationships implicated in complex diseases including cancer.

Over the past ten or more years, many models and algorithms based on machine learning have been developed for analyzing complex biological systems, and contributed to rapid advances in biology and medicine by providing new solutions (Larranaga et al., 2006; McKinney et al., 2006; Fogel, 2008; Upstill-Goddard et al., 2013). Because biological systems are inherently non-linear and dynamic, the proper comprehension of such systems requires interpretive methods that do not rely strictly on linearity and can deal with complex relationships.

Higher-order models represent complex interactions among many factors with higher-order units as their features instead of data variables or linear summations of the variables (Roddick et al., 2008). Higher-order models can more precisely characterize the complicated dependencies embodied in biological phenomena, thus providing better modeling performance than simple linear models (Lehar et al., 2008). Such models based on higher-order representations can be complementary to existing approaches, and can be used to search very large solution spaces efficiently for analyzing complex biological processes including cancer. The range of recent successful applications makes it all the more evident that the need for these models will continue to increase in the near future.

1.2 Problems to be Addressed

Many real-world problems in biological and medical fields require higher-order representation of complex dependency among various factors. Moreover, recent advances in high-throughput molecular techniques have resulted in the exponential growth of the amount of biological data that reflect the interplay between biomolecules on a genome-wide scale. Due to the complexity of the regulatory mechanisms involved and the large number of possible interactions, it is a great need for computational approaches which enable to systematically and efficiently analyze complex biological processes.

This dissertation proposes a higher-order graphical model for dealing with complex relations between many factors and focuses on analyzing co-regulatory genomic interactions on a genome-wide scale for understanding various biological processes including cancers with the new proposed higher-order model. The proposed model can naturally learn the higher-order patterns from high-dimensional data by the process of selecting hyperedges and adjusting their weights. However, since the number of possible hyperedges grows exponentially with the number of features and their combinations, it results in a need for effective learning strategies and suitable model structures to solve complex biological problems. Moreover, it requires the method for extracting meaningful biological knowledge from the learned model.

In this dissertation, we mainly addressed three issues: 1) the advanced model structure for representing higher-order interactions between numerous genomic factors, 2) the improved learning method to efficiently search huge combinatorial feature spaces from very high-dimensional biological data, and 3) the novel method for extracting meaningful biological knowledge from the learned models.

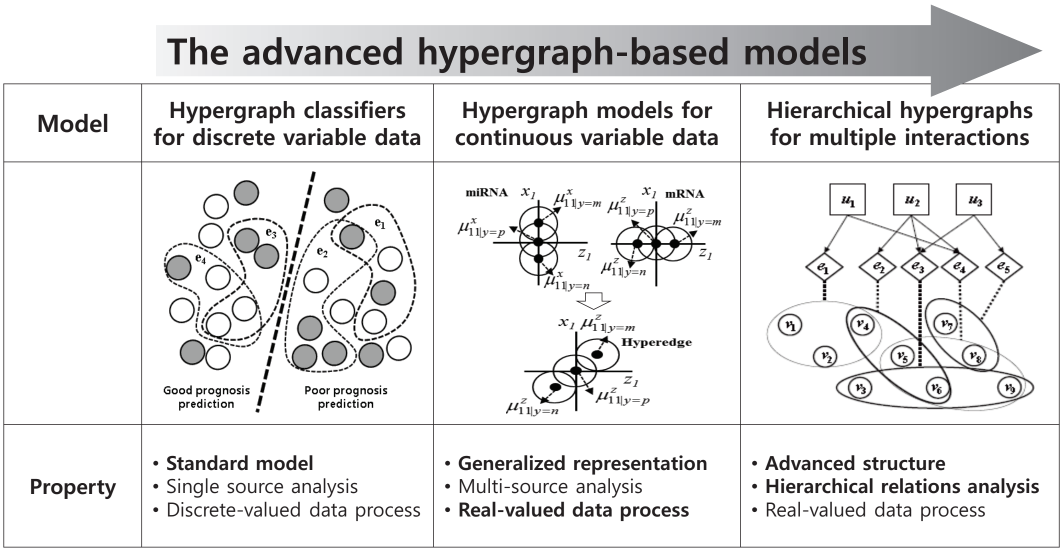
1.3 The Proposed Approach and its Contribution

We propose the advanced class of higher-order graphical models for analyzing complex biological problems incurred by the large number of higher-order interactions, and the improved learning method for efficiently searching huge problem spaces from high-dimensional data. In addition, a novel graph-analyzing method is proposed to extract meaningful biological information and knowledge from the learned models.

The proposed model structure explicitly characterizes higher-order interactions among numerous genomic factors, from which cooperative gene activities in biological processes may be identified. It adopts a flexible hypergraph structure composed of a large population of hyperedges, representing the multi-variable combinations consisting of a variety of genomic factors. Thus, the structure of the proposed model is effective to represent higher-order genomic interactions or complex gene modules for analyzing co-regulatory gene mechanisms in various biological processes including cancer.

The learning of hypergraph models involves searching a huge combinatorial feature space due to its definition and the problem space exponentially enlarges as the number of features increase. This issue becomes more severe when applied to large-scale biological data which consists of several tens of thousands variables. The proposed learning method is able to efficiently search a huge problem space reflecting higher-order relationships between factors by introducing information-theoretical criteria for a guided search into the conventional evolutionary learning approach. The proposed learning method is explained with a sequential Bayesian sampling framework.

Finally, this dissertation proposes a method to enable identify co-regulatory gene modules or to construct gene regulatory networks from the learned higher-order model. Although it is important to extract meaningful information and knowledge from the models in biological and medical fields, the previous studies on hypergraph models focused on the learning efficiency and the model performance rather than knowledge extraction by analyzing the learned model. A network characteriz-



The advanced hypergraph-based models

Model	Hypergraph classifiers for discrete variable data	Hypergraph models for continuous variable data	Hierarchical hypergraphs for multiple interactions
Property	• **Standard model** • Single source analysis • Discrete-valued data process	• **Generalized representation** • Multi-source analysis • **Real-valued data process**	• **Advanced structure** • **Hierarchical relations analysis** • Real-valued data process

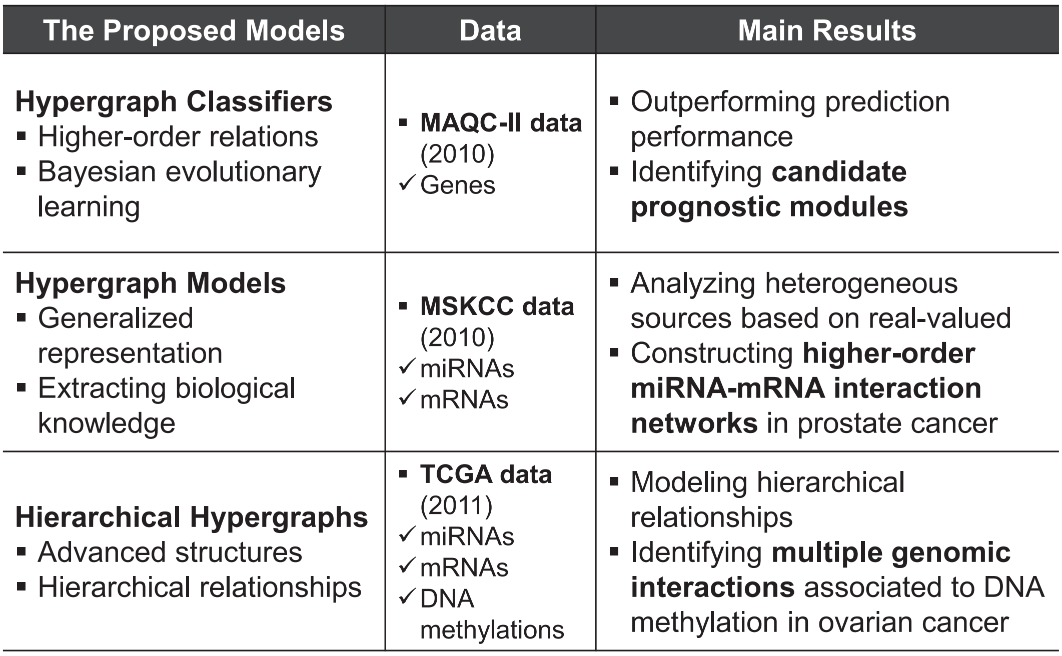
Figure 1.2: The improvement of the proposed models in this dissertation.

ing higher-order genomic interactions is constructed from the leaned hypergraphs based on a minimum-cut approach in this dissertation. Thus, the proposed model can directly extract meaningful knowledge such as co-regulatory gene modules, pathways or networks from various genomic data. Furthermore, it can discover new or potential genomic regulatory circuits which assist our understanding of biological systems including cancer pathogenesis.

Figure 1.2 illustrates the improvement of the proposed models in this dissertation and Figure 1.3 summarized the main results by the proposed model.

1.4 Organization of the Dissertation

This dissertation is organized as follows:



The Proposed Models	Data	Main Results
Hypergraph Classifiers ▪ Higher-order relations ▪ Bayesian evolutionary learning	▪ **MAQC-II data** (2010) ✓Genes	▪ Outperforming prediction performance ▪ Identifying **candidate prognostic modules**
Hypergraph Models ▪ Generalized representation ▪ Extracting biological knowledge	▪ **MSKCC data** (2010) ✓miRNAs ✓mRNAs	▪ Analyzing heterogeneous sources based on real-valued ▪ Constructing **higher-order miRNA-mRNA interaction networks** in prostate cancer
Hierarchical Hypergraphs ▪ Advanced structures ▪ Hierarchical relationships	▪ **TCGA data** (2011) ✓miRNAs ✓mRNAs ✓DNA methylations	▪ Modeling hierarchical relationships ▪ Identifying **multiple genomic interactions** associated to DNA methylation in ovarian cancer

Figure 1.3: Main results by the proposed models

- Chapter 2 presents a survey of the related work. Firstly, we discuss the previous research on the analysis of co-regulatory gene interactions in genomes. Also, we summarize probabilistic graphical models including Bayesian networks, Markov random fields, and hidden Markov models for biological problems. Next, we explain the concept of higher-order model, and summarize previous studies on higher-order graphical models including hypergraphs. In addition, we introduce the applications of hypergraphs and hypergraph-based models in biological problems.

- In Chapter 3, we propose hypergraph classifiers to identify prognostic gene modules for predicting cancer clinical outcomes. The proposed hypergraph classifier is based on evolutionary learning that identifies higher-order gene modules of cancer clinical outcomes. We demonstrate our model can deal with high dimensional data more effectively than state-of-the-art classification models, and identify potential gene modules characterizing prognosis and recurrence risk in cancer.

- Chapter 4 describes the advanced hypergraph model for identifying higher-order genomic interactions from heterogeneous. And we suggest a method for constructing interpretable networks reflecting such higher-order interactions from the learned hypergraph model. We show that the proposed model can build higher-order miRNA-mRNA interaction networks using MSKCC prostate oncogenome data. Also we confirm the biological relevance of the constructed networks through literature review and functional analysis.
- In Chapter 5, we introduce a hierarchical hypergraph model to identify multiple genomic interactions involved in the specific epigenetic mechanisms. A hierarchy are introduced into the hypergraph model by defining two layers. This hierarchical structure allows the proposed model to analyze higher-order genomic relationships at the multi-level regulation. We demonstrate our model can identify higher-order miRNA-mRNA interactions involved in the specific DNA methylation regulation on a genome-wide scale from TCGA data.
- This dissertation is summarized and directions for further research are discussed in Chapter 6.

Chapter 2

Related Work

2.1 Analysis of Co-Regulatory Genomic Interactions from Omics Data

The availability of high-throughput omics data have opened up a new possibility to study the interaction of genetic components underlying the specific biological process such as tumorigenesis at the systems level. Rapid advances in computational approaches which analyze such large-scale data offer a new conceptual framework that can potential revolutionize our view of biology and disease pathologies.

Several years ago, B. Alberts and L. Hartwell noted that biological processes are organized into co-regulatory groups or modules, and that the reductionist approach for studying each process in isolation is limiting (Alberts, 1998; Hartwell et al., 1999). For this reason, one of key issues in current computational systems biology is to systematically analyze gene regulatory mechanisms by using module-based approach from various omics data. Many efforts have taken advantage of this view to investigate a variety of biological processes such as cancer onset, progression and metastasis, which consist of complex interactions among many genetic components on a genome-wide scale (Segal et al., 2001, 2003, 2004, 2005; Bonneau, 2008; Barabási et al., 2011).

Modern cancer research has progressed from identifying biomarkers to systemically exploring gene interactions (Hornberg et al., 2006; Wang et al., 2007; Liu et al., 2012). Many studies have attempted to describe how genetic components interact on the system level. Computational methods, which analyze gene regulatory networks and interactions on a genome-wide scale from high-throughput biological data, have flourished in recent decades (Bar-Joseph et al., 2003; Schlitt and Brazma, 2007; Yan et al., 2007; Lee and Tzou, 2009; Joung et al., 2012; Mitra et al., 2013). In addition, systems biology approaches to study miRNA regulation were designed to understand the development of multiple human malignancies (Kim et al., 2006; Bandyopadhyay et al., 2010; Volinia et al., 2010). Moreover, recent studies have focused on reconstructing regulatory networks by integrating miRNAs and other molecules such as mRNAs, transcriptional factors, and proteins for different physiological and pathological conditions (Shalgi et al., 2007; Bonnet et al., 2010a; Nasser et al., 2010; Li et al., 2011; Lu et al., 2010; Zhang et al., 2011).

Those approaches have helped to simplify complex biological mechanisms by systemically analyzing the relationships between genetic elements at the genome level. However, many studies on this issue use an approach considering relationships between only two factors for analyzing the interactions among genes. In addition, we are still far from understanding the mechanisms of cooperative regulations among various components in a specific biological process. Therefore, inferring regulatory networks by taking into consideration the complex dependencies among genetic factors remains a formidable challenge.

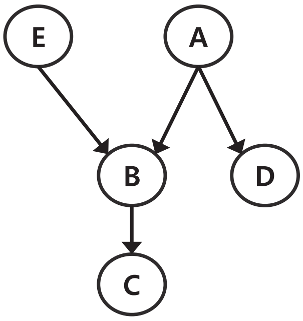
2.2 Probabilistic Graphical Models for Biological Problems

Probabilistic graphical models (PGMs) have been applied to many real-world problems. The general framework of PGMs uses ideas from discrete data structures in computer science to efficiently encode and manipulate probability distributions over high-dimensional spaces, often involving hundreds or even many thousands of variables (Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012). These models have been used in an enormous range of application domains, which include: medical diagnosis, biological network reconstruction, speech recognition, natural language processing, intelligent control, and many more.

2.2.1 Bayesian Networks

A Bayesian network (Heckerman et al., 1995; Jensen, 1996; Friedman et al., 1997; Neapolitan, 2004; Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012) is a graphical model that encodes probabilistic relationships among variables of interest. This model is more suitable for analyzing biological data because it can represent cause and effect relationships. Graphical model has several advantages for data analysis when used in conjunction with Bayesian statistical techniques. Firstly, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Second, Bayesian networks can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Third, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data.

Bayesian networks compactly represent the joint probability distribution over a set of random variables via a directed acyclic graph (DAG). In the framework of probabilistic graphical model, the concept of conditional independence is exploited



Conditional independence statements:
I(A; E), I(B; D | A, E), I(C; A, D, E | B),
I(D; B, C, E | A), and I (E; A, D)

Product form of joint distribution:
P(A, B, C, D, E) = P(A)P(B|A, E)P(C|B)P(D|A)P(E)

Figure 2.1: An example of a simple Bayesian network structure

for efficient representation of joint probability distribution. For three variable sets $\{X, Y, Z\}$, X is conditionally independent from Y given the value of Z. The Bayesian network structure encodes various conditional independencies among the variables as follows. A Bayesian network assumes a directed acyclic graph structure where each node corresponds to a variable and an edge is a direct probabilistic dependency between the two connected nodes. Formally, the DAG structure asserts that each node is independent of all its non-descendants conditioned on its parent nodes. A Bayesian network consisting of n variables, $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, represents a joint probability distribution $\mathbf{P}(\mathbf{X})$ given by the product of all conditional probability:

$$\mathbf{P}(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | pa^G(X_i)) \tag{2.1}$$

where $pa^G(X_i)$ is the set of parents of X_i in the DAG structure G, and $\mathbf{P}(\mathbf{X})$ reflects the properties of Bayesian network. A graph G specifies a product form as in Figure 2.1. To fully specify a joint distribution, we also need to specify each of the conditional probabilities in the product form. The second part of the Bayesian network describes these conditional distributions, $P(X_i | pa^G(X_i))$ for each variable

X_i.

Bayesian network structure learned from data can provide us insight into the complicated cause and effect relationship among a set of variables. Thus, it is applicable for extracting knowledge from data. As such, Bayesian network has been widely applied in various areas including cancer diagnosis (Nikovski, 2000; Gevaert et al., 2006; Cruz-Ramírez et al., 2007) and gene expression analysis (Friedman et al., 2000; Segal et al., 2003; Imoto et al., 2004; Liu et al., 2013).

2.2.2 Markov Random Fields

A Markov random field (MRF) (Kindermann and Snell, 1980; Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012) also known as a Markov network, or a probabilistic independence network, is a set of random variables having a Markov property described by an undirected graph (Figure 2.2 (a)). Every variable X_i is represented by a node in the graph and the nodes are connected by undirected edges. Let $adj(X_i)$ be all the nodes that are adjacent (i.e., directly connected) to X_i, then the edges in a Markov field are places in such way that:

$$\forall X_j \in \chi \backslash X_i \cup adj(X_i); X_i \perp\!\!\!\perp X_j | adj(X_i) \tag{2.2}$$

where χ is the value set of X_i and $adj(X_i)$ acts as the Markov blanket of X_i. In contrast to belief networks, there are no conditional probability functions connected to nodes. Instead, each *clique* in the graph is provided with a *potential* $\psi_c(\cdot)$ which assigns a non-negative real value to all combinations of values of nodes. A clique is defined as a subset of the nodes in a graph such that there exists a link between all pairs of nodes in the subset. In other words, the set of nodes in a clique is fully connected. Furthermore, a *maximal clique* is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.

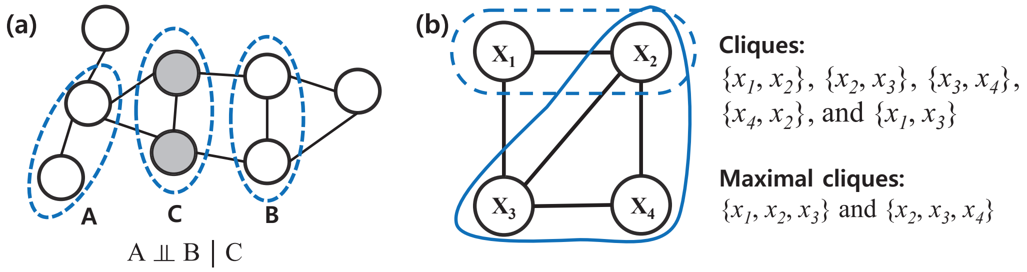

(a)
A
C
B
A ⫫ B | C
(b)
X_1
X_2
X_3
X_4
Cliques:
$\{x_1, x_2\}$, $\{x_2, x_3\}$, $\{x_3, x_4\}$,
$\{x_4, x_2\}$, and $\{x_1, x_3\}$
Maximal cliques:
$\{x_1, x_2, x_3\}$ and $\{x_2, x_3, x_4\}$

Figure 2.2: (a) An example of an undirected graph in which every path from any node in set A to any node in set B passes through at least one node in set C. Consequently the conditional independence property $A \perp\!\!\!\perp B|C$ holds for any probability distribution described by this graph. (b) A four-node undirected graph showing a clique (straight line) and a maximal clique (dotted line).

These concepts are illustrated by the undirected graph over four variables shown in Figure 2.2 (b). Let us denote a clique by C and the set of variables in that clique by $\mathbf{x}_C$. Then the joint distribution is written as a product of *potential function* $\psi_C(\mathbf{x}_C)$ over the maximal cliques of the graph

$$P(\mathbf{x}) = \frac{1}{Z}\prod_C \psi_C(\mathbf{x}_C), \tag{2.3}$$

where the quantity Z, called the partition function, is a normalization constant and is given by

$$Z = \sum_{\mathrm{x}}\prod_C \psi_C(\mathbf{x}_C) \tag{2.4}$$

which ensures that the distribution $p(x)$ given by (2.4) is correctly normalized.

In this manner, MRFs can represent circular dependencies between variables and it is useful in the cases in which direction of influence has no meaning, for example when variables represent pixels in image or atom in a protein molecule. As such, MRFs have seen wide application in many areas, including computer vision (Li, 1995, 2009; Wang et al., 2013), and bioinformatics (Demirkaya et al., 2005; Wei and Li, 2007; Chen et al., 2011).

2.2.3 Hidden Markov Models

A hidden Markov model (HMM) (Rabiner and Juang, 1986; Eddy, 1996; Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012) can be viewed as a specific instance of the state space model of Figure 2.3 in which the latent variables are discrete. The HMM models a sequence of observations $X = \{x_t\}_{t=1}^{T}$ by assuming that there is an underlying sequence of states $Y = \{y_t\}_{t=1}^{T}$ drawn from a finite state set S. To model the joint distribution $p(\mathbf{y}, \mathbf{x})$ tractably, an HMM makes two independence assumptions. First, it assumes that each state depends only on its immediate predecessor, that is, each state yt is independent of all its ancestors $y_1, y_2, ..., y_{t-2}$ given its previous state y_{t-1}. Second, an HMM assumes that each observation variable x_t depends only on the current state y_t. With these assumptions, we can specify an HMM using three probability distributions: first, the distribution $p(y_1)$ over initial states; second, the transition distribution $p(y_t|y_{t-1})$; and finally, the observation distribution $p(x_t|y_t)$. That is, the joint probability of a state sequence $\mathbf{y}$ and an observation sequence $\mathbf{x}$ factorizes as

$$p(\mathrm{y}, \mathrm{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t), \tag{2.5}$$

where to simplify notation, we write the initial state distribution $p(y_1)$ as $p(y_1|y_0)$.

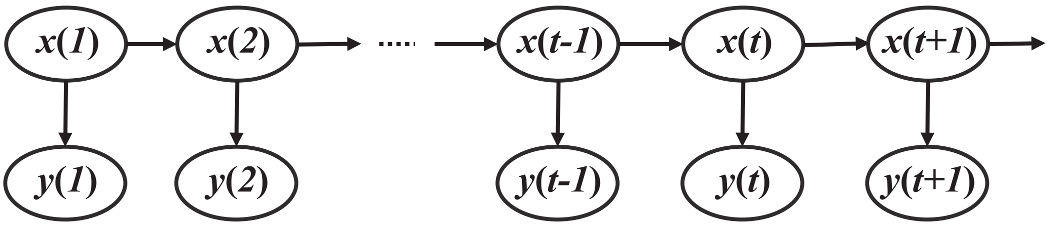


Figure 2.3: Graphical structure of hidden Markov model. We can represent sequential data using a Markov chain of latent variables, with each observation conditioned on the state of the corresponding latent variable.

The HMM is widely used in speech recognition (Rabiner, 1989), natural language modeling (Manning and Schütze, 1999), and for the analysis of biological data (Eddy, 1998; Krogh et al., 2001; Wheeler et al., 2013; Bonneville and Jin, 2013).

2.3 Higher-order Graphical Models for Biological Problems

2.3.1 Higher-Order Models

We generally assume pairwise relationships among the objects of our interest in machine learning problem setting. An object set endowed with pairwise relations can be naturally described as a graph, in which the vertices represent the objects, and any two vertices that have some kind of relationship are joined together by an edge. However, in many real-world problems, relationships among the objects of our interest are more higher-order than pairwise, and thus representing a set of their complex relationships as general undirected or directed graphs is not complete.

A higher-order model uses higher-order units as features. While linear models are difficult to reflect high order dependency embodied in the data, higher-order models can represent higher-order relationships, thus fitting the complex solution spaces including nonlinearity.

A higher-order unit can be defined to a feature represented with patterns or function values derived from raw attributes of given data (Roddick et al., 2008; Lehar et al., 2008). When A is a subset of $\mathbf{X}$=$\{x_1,\cdots,x_m\}$, the set of attributes of the given data, formally, a feature **f** is defined as follows:

$$f = g(A) \tag{2.6}$$

where g denotes an arbitrary function. When f is an identity function, **f** denotes a raw attribute. Also, f is a linear feature with weighted summation of the elements of A in case of $g = \sum_{x_i \in A} w_i x_i$. On the other hands, **f** becomes a higher-order feature when g is a function with two or more order such as sin, exp, and $\prod_{x_i \in A} x_i$.

In this dissertation, we use an individual represented with a conjunction of attribute values of data, and a population of them as a higher-order unit and a higher-order model, respectively. This conjunction-based individual representation enhances the interpretability of the models more compared with units based on numerical functions. Also, the individual and the population in our study can be represented with a hyperedge and a hypergraph.

2.3.2 Hypergraphs

A hypergraph (Berge, 1989; Gallo et al., 1993; Zhou et al., 2007) is a generalized graph for representing complex interactions. In the hypergraph construct, the edge in a conventional graph (which connects two vertices) is generalized to the hyperedge, which connects more than two vertices concurrently. A hyperedge is weighted by the strength of the higher-order dependency among its elements.

Unlike conventional graphs, hypergraphs are suitable for explicitly representing higher-order relationships among many features.

Formally, a hypergraph H is formulated as a triple $H = \{V, E, W\}$, where V, E, and W denote the sets of vertices v, hyperedges e, and hyperedge weights $w(e)$, respectively. A hyperedge of weight $w(e)$ is represented as a subset of V. Let $d(v)$ and $\delta(e)$ denote the degree of a vertex v and the degree of a hyperedge e, respectively. Each degree is then defined as follows:

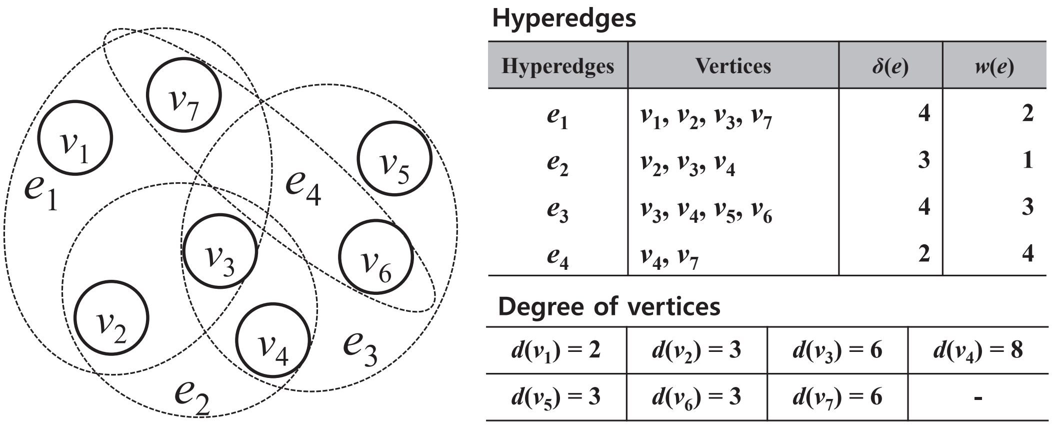
$$d(v) = \sum\nolimits_{e \in E} w(e)h(v, e) \tag{2.7}$$

and

$$\delta(e) = |e| \tag{2.8}$$

where |e| is the cardinality (number of vertices) of e and $h(v, e)$ is an indicator function that returns 1 if v is an element of e and 0 otherwise. A hyperedge with degree k is called a k-hyperedge and a hypergraph consisting solely of k-hyperedges is a k-hypergraph. Figure 2.4 shows an example of a hypergraph. A high-degree vertex can be regarded as a hub of the hypergraph structure, which may signify an informative feature for classifying the given data. Moreover, a hyperedge with higher degree embodies more specific information, whereas one with lower degree characterizes more general patterns. Because they naturally represent higher-order interactions, hypergraphs have become a popular choice for solving a range of problems (Hu et al., 2008; Klamt et al., 2009; Bu et al., 2010).

According to its applications, vertices and hyperedges denote different objects. In Zhou's study (Zhou et al., 2007), a vertex is a data instance and a hyperedge denotes a set of vertices with identical attribute values. In Kok's study to build Markov logic networks from relational database, on the other hands, a vertex represents



Hyperedges

Hyperedges	Vertices	$\delta(e)$	$w(e)$
e_1	v_1, v_2, v_3, v_7	4	2
e_2	v_2, v_3, v_4	3	1
e_3	v_3, v_4, v_5, v_6	4	3
e_4	v_4, v_7	2	4

Degree of vertices

$d(v_1) = 2$	$d(v_2) = 3$	$d(v_3) = 6$	$d(v_4) = 8$
$d(v_5) = 3$	$d(v_6) = 3$	$d(v_7) = 6$	-

Figure 2.4: An example of hyperedges in a hypergraph

a discrete data attribute (variable) and a hyperedge means logical relation among them (Kok and Domingos, 2009). Same to Kok's representation, a vertex means discrete variable value and a hyperedge represents an arbitrary combination of vertices in Zhang's models (Zhang, 2008).

The understanding of complex biological systems is a fundamental issue in computational biology. In particular, when analyzing topological properties of biological networks, one often tends to substitute the term "network" for "graph", or uses both terms interchangeably. From a mathematical perspective, this is not fully correct, because many functional relationships in biological networks are more complicated than what can be represented in graphs.

As mentioned above, graphs are combinatorial models for representing relationships (edges) between certain objects (vertices or nodes). In biology, the vertices typically illustrate genes, transcription factors, proteins, metabolites, or other biological components, whereas the edges represent functional relationships or inter-

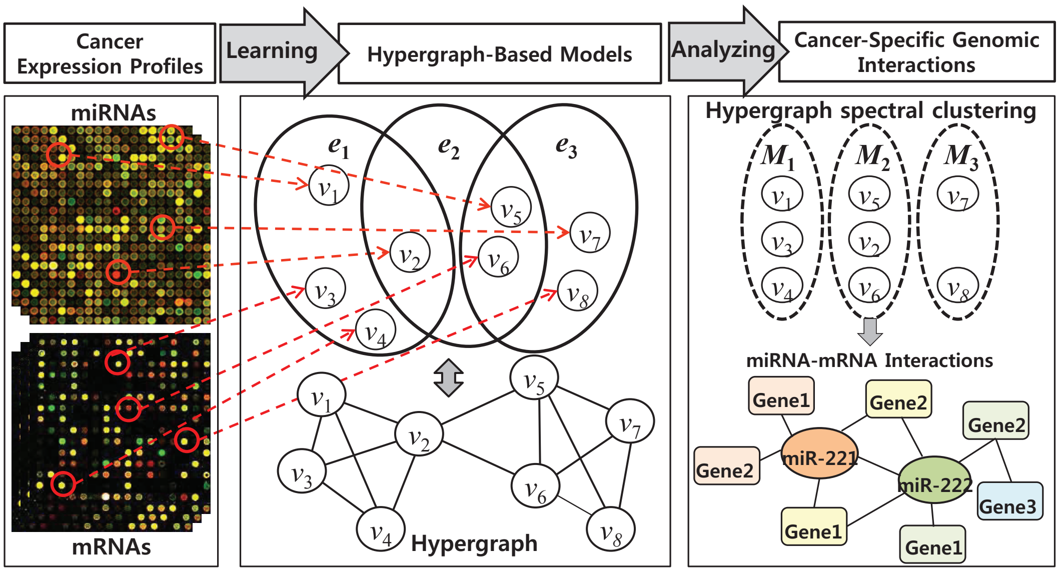


Figure 2.5: Modeling genomic interactions via hypergraph-based models

actions between the vertices such as “binds to”, “regulates to”, or “is converted to”. A key property of graphs is that every edge connects two vertices. However, many biological processes including cancer are characterized by more than two participating cooperators and are thus not bilateral. Hence, multilateral relations are not compatible with general graph edges. In addition, transformation to a graph representation is usually possible but may imply a loss of information that can lead to wrong interpretations subsequently.

Hypergraphs provide a framework that helps to overcome such conceptual limitations. As the name indicates, a hypergraph is a generalized graph by allowing edges to connect more than two vertices, which may facilitate a more precise representation of higher-order interactions in biological processes. Thus, hypergraph-based models are suitable for representing a knowledge network to investigate complex biological phenomena and they have been successfully used for diverse biological problems (Ha et al., 2007; Kim et al., 2007; Tian et al., 2008; Klamt et al., 2009; Kim et al., 2010). Figure 2.5 shows an example of a hypergraph-based models for modeling cancer-specific genomic interactions from cancer expression profiles.

Chapter 3

Hypergraph Classifiers for Identifying Prognostic Modules in Cancer

3.1 Overview

Predicting the clinical outcomes of cancer patients is a challenging task in biomedicine. A personalized and refined therapy based on predicting prognostic outcomes of cancer patients has been actively sought in the past decade. Accurate prognostic prediction requires higher-order representations of complex dependencies among genetic factors. However, identifying the co-regulatory roles and functional effects of genetic interactions on cancer prognosis is hindered by the complexity of the interactions.

In this chapter, we introduce a new population-based model that uses an evolutionary learning method to predict clinical outcomes of cancer patients (Figure 3.1) (Kim et al., 2013a). The model handles complex genomic interactions by means

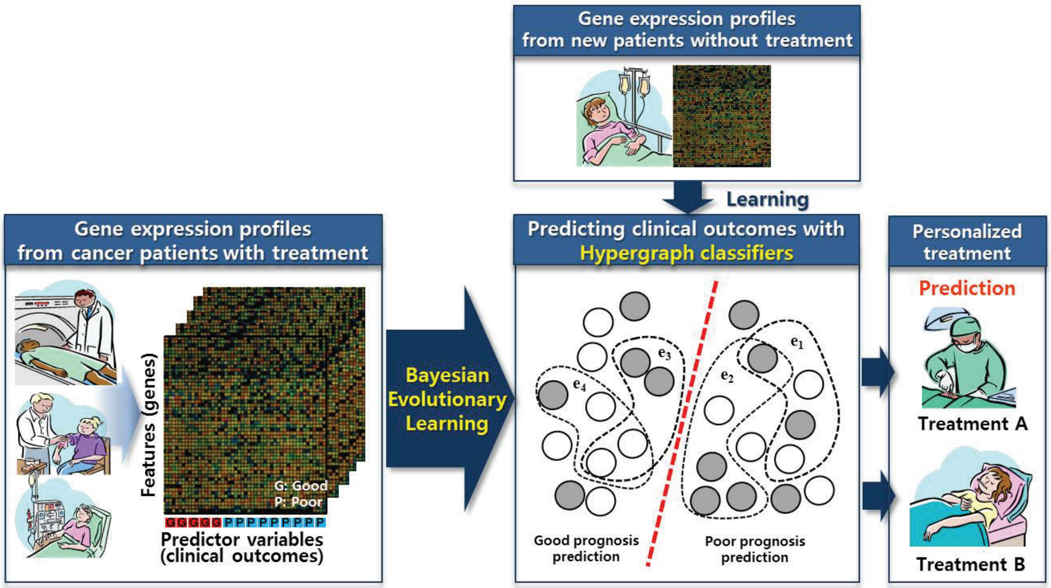

Gene expression profiles from new patients without treatment
Learning
Gene expression profiles from cancer patients with treatment
Features (genes)
G: Good
P: Poor
GGGGGPPPPPPPPPP
Predictor variables (clinical outcomes)
Bayesian Evolutionary Learning
Predicting clinical outcomes with Hypergraph classifiers
e1
e2
e3
e4
Good prognosis prediction
Poor prognosis prediction
Personalized treatment
Prediction
Treatment A
Treatment B

Figure 3.1: Overview of the hypergraph classifier based on Bayesian evolutionary learning for predicting cancer clinical outcomes from cancer genomic data

of a flexible hypergraph structure comprising a large population of hyperedges, representing the multi-variable combinations corresponding to all potential genes or markers. Each hyperedge is weighted by its discriminative ability to predict prognostic outcomes. Thus, each hyperedge potentially behaves as a prognostic module influencing the cancer clinical outcomes.

The model learning involves the search of a high-dimensional space reflecting the higher-order relationships between factors. To learn the model from a dataset comprising several tens of thousands of genetic variables, an evolutionary method based on sequential Bayesian sampling scheme is applied (Ha et al., 2013). The proposed Bayesian evolutionary algorithm is designed upon a standard evolutionary computation framework. Variation, evaluation, and selection are repeated as a sequential Bayesian sampling process, where the posterior distribution is recursively calculated from the prior distribution by estimating the likelihood from fitness measurements. Using this Bayesian formulation of evolutionary computation, the model can determine the problem-specific bias as a guideline for efficient search of a huge combinatorial feature space. This study adopts an information theoretic co-regulatory measure called mutual information, and the model complexity for the distribution. The information theoretic measure enhances the efficiency of the evolutionary search, while the complexity retains a compact model size by controlling the parsimony.

The proposed model is evaluated on MAQC-II breast cancer and multiple myeloma gene expression data (Shi et al., 2010). The proposed model demonstrates high classification performance for predicting prognosis in patients, and can identify higher-order prognostic biomarkers of cancer clinical outcomes. Moreover, our model directly identifies potential modules of informative genes that characterize prognosis and recurrence risk in cancer.

3.2 Analyzing Gene Modules for Cancer Prognosis Prediction

Prognostic prediction is an important task in clinical medicine. Estimating the clinical outcomes of patients and the potential effects of treatment is crucial. A refined treatment based on likely clinical outcomes is especially necessary in oncology, because cancer progression varies between patients. By accurately estimating the clinical response to treatment, clinicians can personalize and hence provide an improved therapy for a patient.

Gene expression profiling has been widely used to identify tumor heterogeneity, and has led to the discovery of molecular signatures of potential prognostic and therapeutic interest (Simon, 2003; Fan et al., 2010; Goodison et al., 2010). As such, it is recognized as a powerful source for improving prognostic assessment and treatment selection in cancer medicine. Moreover, cancer prognosis is associated with combinatorial and modular regulation by multiple genetic factors. Thus, for more precise prediction of cancer clinical outcomes, the higher-order relationships among genetic factors must be deduced from gene expression profiles. However, the complexity of gene interactions renders this task extremely challenging.

Predictive methods, which classify patient outcomes on a genome-wide scale from high-throughput biological data, have flourished in recent decades. Many studies have adopted computational approaches, such as machine learning-based models (Veer et al., 2002; Street et al., 1995; Koziol et al., 2009; Sun et al., 2011; Verduijn et al., 2007; Gevaert et al., 2006; Han et al., 2011; Berchuck et al., 2005; Kim et al., 2012a) and statistical methods (Braitman and Davidoff, 1996; Huang et al., 2003; Boulesteix et al., 2008; Matsui et al., 2007), to predict prognosis from cancer genomic data. However, few of the existing approaches address the higher-order interactions between genes involved in cancer prognosis.

Predicting outcomes from higher-order gene relationships requires searching of an exponential search space consisting of tens of thousands of genes. Such a huge combinatorial feature space cannot be exhaustively searched using a gradient method, and is instead undertaken by various feature selection methods (Saeys et al., 2007). Typically, these approaches reduce the problem space by individually evaluating each gene, assuming independence between features. However, such restrictions may not capture the important genes involved in higher-order relationships underlying pathological processes.

3.3 Hypergraph Classifiers for Identifying Cancer Gene Modules

3.3.1 Hypergraph Classifiers

The proposed population-based model uses hypergraph structures composed of a large collection of hyperedges playing the role of a weak classifier. These hyperedge ensembles are called hypergraph classifiers. The unlabeled data can be predicted by assembling this population of many weak classifiers.

We assume that in the n-th data instance, a set of class labels denoting clinical outcome, Y, and a hypergraph, H, are given. The y value whose weighted sum of hyperedges corresponding to the genetic variables in $\mathbf{x}^{(n)}$ is the largest among the elements of Y is called the class label of $\mathbf{x}^{(n)}$, denoted $\hat{y}^{(n)}$. Specifically, the class label is determined as follows:

1. Calculate c_y, the sum of weights for $y \in$Y over all hyperedges in E:

$$c_y = \sum_{i=1}^{|E|} \left\{ w(e_i) f(\mathbf{x}^{(n)}, e_i) \varphi(y^{(n)}, y_i) \right\} \tag{3.1}$$

   where $|E|$ denotes the hyperedge set and $w(e_i)$ is the weight of e_i.

2. Predict the class label of $\mathbf{x}^{(n)}$, $\hat{y}^{(n)}$, as the y value with the highest total weight:

$$\hat{y}^{(n)} = \arg\max_{y \in Y} c_y \tag{3.2}$$

In Equations (3.1) and (3.2) above, $f(\mathbf{x}^{(n)}, e_i)$ and $\varphi(y^{(n)}, y_i)$ denote the matching and indicator functions, which return 1 if e_i matches $\mathbf{x}^{(n)}$ and if $y^{(n)} = y_i$, respectively. These functions are defined as Equations (3.14) and (3.15) in the next subsection. This classification process is similar to an learning classifier system (LCS) (Holland, 1980), in which each classifier participates in classifying the unlabeled data as a significant condition-action rule. However, the proposed hypergraph classifier focuses on the model structure (the entire connected ensemble of hyperedges), rather than on each hyperedge. The hyperedges composing the population exert the main influence on the classification performance. In the next subsection, we explain how the population is generated and how the model is learned by an evolutionary method.

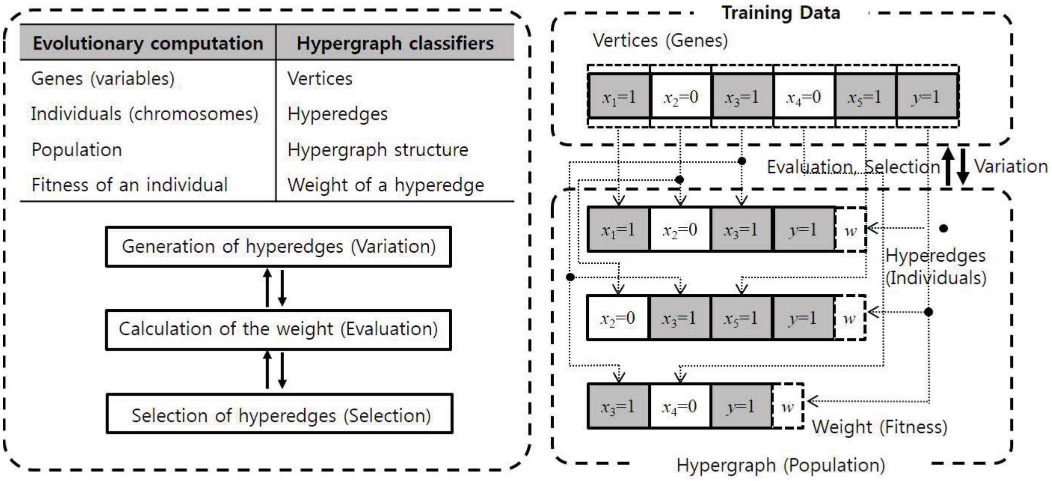
3.3.2 Bayesian Evolutionary Algorithm

The Bayesian evolutionary algorithm implements an evolutionary learning method based on sequential Bayesian sampling. A standard evolutionary computation process that iterates the generation of individuals (variation), calculation of the fitness (evaluation), and selection of individuals (selection) is implemented with the Bayesian sampling framework where the posterior distribution is recursively computed by estimating the likelihood from the prior distribution. Figure 3.2 presents the terms of hypergraph classifiers and their corresponding terms in standard evolutionary computation schema. A naive evolutionary method may be inefficient when the problem involves the searching of vast and complex solution spaces. However, Bayesian evolutionary algorithm can efficiently search the space by introducing problem-specific knowledge to the prior distribution.

Let H_t be a population at the t-th generation. For a dataset $D = (X, Y)$, where $X = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ and $Y = \{y^{(n)}\}_{n=1}^{N}$ are given, Bayes' rule specifies the posterior distribution of H_t as the conditional probability:

$$p(H_t|X, Y) = \frac{p(Y|X, H_t)p(H_t|X)}{p(Y|X)} \tag{3.3}$$

where $p(Y|X, H_t)$ and $p(H_t|X)$ denote the likelihood and the prior, respectively. In Equation (3.3), $p(Y|X)$ is a normalizing constant because it is independent of H_t.



Evolutionary computation	Hypergraph classifiers
Genes (variables)	Vertices
Individuals (chromosomes)	Hyperedges
Population	Hypergraph structure
Fitness of an individual	Weight of a hyperedge

Generation of hyperedges (Variation)
Calculation of the weight (Evaluation)
Selection of hyperedges (Selection)
Training Data
Vertices (Genes)
x_1=1 x_2=0 x_3=1 x_4=0 x_5=1 y=1
Evaluation, Selection
Variation
x_1=1 x_2=0 x_3=1 y=1 w
Hyperedges (Individuals)
x_2=0 x_3=1 x_5=1 y=1 w
x_3=1 x_4=0 y=1 w
Weight (Fitness)
Hypergraph (Population)

Figure 3.2: Hypergraph classifiers in standard evolutionary computation

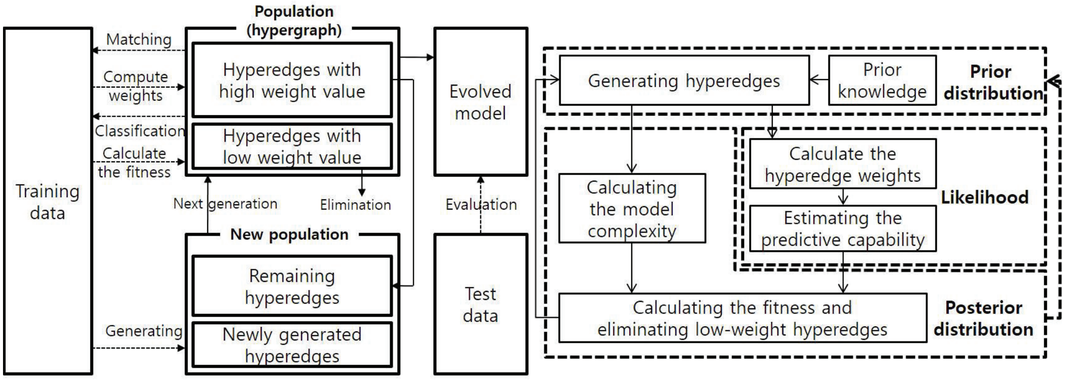
Thus, the posteriori distribution is proportional to the product of the likelihood and the prior:

$$p(H_t|X,Y) \propto p(Y|X,H_t)p(H_t|X) \tag{3.4}$$

The aim of the evolutionary process is to maximize the model fitness F_t, defined as the logarithm of the posterior:

$$\begin{aligned} F_t &= \log p(Y|X,H_t) + \log p(H_t|X) \ , && (3.5) \\ H^* &= \arg\max_{H_t} F_t && (3.6) \end{aligned}$$

Finally, the evolution of hypergraph classifiers is regarded as the maximum a posteriori (MAP) process in the Bayesian learning. Figure 3.3 describes the evolving process of hypergraph classifiers learned by the Bayesian approach.



(a) Evolving flow of hypergraph classifiers (b) Bayesian view of evolving hypergraph classifiers

Figure 3.3: Flowchart of Bayesian evolutionary learning of hypergraph classifiers

3.3.3 Bayesian Evolutionary Learning for Hypergraph Classifiers

The model fitness is computed from the prior and the likelihood. The empirical prior distribution $p(H_t|X)$ can be defined by prior knowledge of the problem, which enhances the efficiency of the evolutionary search. In this study, the prior includes two factors. One is mutual information (MI) between each variable and the class label, specified as the relationships between the data rather than a uniform distribution. MI is an information-theoretic measure that specifies the degree of conditional independence between two random variables. Here, it is used as a co-regulatory measuring criterion for efficiently selecting genes for hyperedge generation. The other factor is the model complexity. The prior is defined to prefer the most parsimonious, or smallest, model. This prior not only ensures that genetic variables relevant to prognostic outcomes are more frequently included in the model, but also retains the model compact. Therefore, the current empirical prior for generating hyperedges is calculated from the MIs and the previous posterior, $p(H_{t-1}|Y, X)$

reflecting the model complexity $|H_{t-1}|$:

$$p(H_t|X) \propto p(H_{t-1}|Y, X) \tag{3.7}$$

$$p(H_t|X) \propto \frac{1}{|H_{t-1}|} \prod_{e \in E_{new}} P(e) \approx \frac{1}{|H_{t-1}|} \prod_{e \in E_{new}} \prod_{x_i \in e} P_I(x_i) \tag{3.8}$$

s.t.

$$P_I(X_i) = \frac{I(X_i;Y)^\eta}{\sum_{j=1}^{|X|} \{I(X_i;Y)^\eta\}}, \ |H_t| = \sum\nolimits_{e \in E_t} \delta(e), \text{ and } E_{new} = E_t - E_{t-1},$$

where E_t is the hyperedge set of H_t, and $P(e)$ denotes the probability with which a hyperedge e is generated. $P_I(X_i)$ indicates the probability of selecting the i-th genetic factor X_i, which depends on the MI between X_i and the class label Y, $I(X_i;Y)$. The nonnegative constant η regulates the influence of MIs on the gene selection. The prior distribution influences hyperedge construction in every generation. Specifically, a hyperedge is generated as follows:

1. Select the data instance from which to subsample a hyperedge.
2. Probabilistically determine the degree of the hyperedge within a predefined range:

$$P(\delta(e) = K) = \frac{|E^K_{t-1}| + \varepsilon}{\sum_{k=K_{\min}}^{K_{\max}} \left(|E^k_{t-1}| + \varepsilon\right)} \tag{3.9}$$

where E^k_{t-1} denotes a set of k-hyperedges at generation t-1 and ε is a smoothing constant.

3. Probabilistically select the variables based on $P_I(X_i)$.
4. Construct a hyperedge from a set of variable values and the class label of the selected data instance.
5. Add the generated hyperedges to the population.

Hyperedge generation in our model differs from that of LCS, where each classifier is generated by genetic operations such as crossover and mutation. Our model can efficiently search a high-dimensional space without a heavy computational cost, because it guarantees that a pattern in a hyperedge always exists in the training data.

The likelihood is defined to represent the discriminative capability of the model. To achieve this, we assume that the capability grows by increasing the difference of the weighted sum between the correctly and incorrectly matched hyperedges for all training data. A hyperedge is said to be correctly matched if it matches a given data instance and the label of the hyperedge equals that of the instance. On the other hand, an incorrectly matched hyperedge is matched to an instance with a different class label than itself. Since the instances are independent, the likelihood is estimated as the product of the empirical likelihoods on the given data:

$$p(Y|X, H_t) = \prod_{n=1}^{N} p(y^{(n)}|x^{(n)}, H_t) \tag{3.10}$$

and the empirical likelihood is defined by:

$$p\left(y^{(n)}|x^{(n)}, H_t\right) \equiv \frac{\sum_{i=1}^{|E_t|} w(e_i)\left\{f_i^{(n)} \cdot \varphi_i^{(n)} - f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)}\right)\right\}}{\sum_{i=1}^{|E_t|} w(e_i)} \tag{3.11}$$

where $w(e_i)$ denotes the weight of the i-th hyperedge. Thus, we have

$$p(Y|X,H_t) = \prod_{n=1}^{N} \left[\frac{\sum_{i=1}^{|E_t|} w(e_i) \left\{ f_i^{(n)} \cdot \left(2\varphi_i^{(n)} - 1 \right) \right\}}{\sum_{i=1}^{|E_t|} w(e_i)} \right] \tag{3.12}$$

$$\text{s.t. } f_i^{(n)} = f(x^{(n)}, e_i) \text{and} \varphi_i^{(n)} = \varphi(y^{(n)}, y_i) \tag{3.13}$$

with the matching and indicator functions respectively defined as follows:

$$f_i^{(n)} = f(x^{(n)}, e_i) = \begin{cases} 1, \text{if} \exp\left\{ c(x^{(n)}, e_i) - \left| e_i - \{y_i\} \right| \right\} > \theta \\ 0, \text{otherwise} \end{cases}, \tag{3.14}$$

$$\varphi_i^{(n)} = \varphi(y^{(n)}, y_i) = \begin{cases} 1, \text{if} y^{(n)} {=} y_i \\ 0, \text{otherwise} \end{cases}, \tag{3.15}$$

where $c(x^{(n)}, e_i)$ is the matching number, defined as the number of hyperedge variables that equal their corresponding variables in $\mathbf{x}(n)$. The matching threshold θ smoothes and enhances robustness against data noise by allowing partial matching. Also, $f_i^{(n)} \cdot \varphi_i^{(n)}$ and $f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)}\right)$ equal 1 for a correctly and incorrectly matched hyperedge, respectively, and 0 otherwise. The weight of a hyperedge is a function of correctly and incorrectly matched cases:

$$\begin{aligned} w(e_i) &= \left| y_i \right|^{\beta} \left\{ \alpha \sum_{n=1}^{N} f_i^{(n)} \cdot \varphi_i^{(n)} - (1-\alpha) \sum_{n=1}^{N} f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)} \right) \right\} && (3.16) \\ &= \left| y_i \right|^{\beta} \left\{ \sum_{n=1}^{N} f_i^{(n)} \cdot \varphi_i^{(n)} - (1-\alpha) \sum_{n=1}^{N} f_i^{(n)} \right\}, && (3.17) \end{aligned}$$

where α is a constant for preferring more correct or less incorrect predictions. For data whose class labels are imbalanced, a quantity $|y_i|$, denoting the number of data with class label y_i, and a negative constant β, are introduced into the weight function. If $w(e)$ is negative, it is reset to zero to prevent the construction of a negatively weighted graph. The model fitness is then reformulated from (3.5) using the defined prior (3.8) and the estimated likelihood (3.13):

$$F_t = \log p(Y|X, H_t) + \log p(H_t|X)$$

$$\approx \sum_{n=1}^{N} \log \frac{\sum_{i=1}^{|E_t|} \left\{ w(e_i) f_i^{(n)} \left(2\varphi_i^{(n)} - 1 \right) \right\}}{\sum_{i=1}^{|E_t|} w(e_i)} + \lambda \left| H_t \right| + \zeta \sum_{e \in E_t} \log \sum_{x_i \in e} P_I(x_i) \tag{3.18}$$

where λ and ζ denote a negative constant for regularizing the model size and a positive value for regulating the selection power of the variables in the prior, respectively. To increase the fitness, hyperedges with high weight survive at every generation; in addition, a hyperedge is generated from variables with large $P_I(x)$, and the proportion of lower-degree hyperedges is increased.

As the population changes, the hypergraph structure evolves by removing hyperedges with relatively low weight and replacing them with new hyperedges at every generation. To prevent the removal of highly discriminating hyperedges, the number of replaced hyperedges decreases to a specific value as the iterations proceed. The number of replacements at the t-th generation is adaptively determined:

$$R_t = \frac{R_{max} - R_{min}}{\exp(t/\kappa)} + R_{min} \tag{3.19}$$

where t is the iteration number of the learning process, and R_{max} and R_{min} denote the maximum and minimum boundary values of R_t, respectively. κ is a constant that moderates the speed at which the system proceeds from R_{max} to R_{min}.

3.4 Predicting Cancer Clinical Outcomes Based on Gene Modules

3.4.1 Data and Experimental Settings

The gene expression data have been widely used in various applications. They include diagnosis, early detection, monitoring of disease progression, risk assessment, prognosis, complex medical product characterization and prediction of response to treatment. For this reason, many classification models for microarray data have been proposed for being applied to the biological and medical fields. Herein, the published benchmarking studies on classifiers for microarray data have split data into two sets: a dataset used for training and the other set for validation, with randomness. This design assumes that the training and validation sets are produced by unbiased sampling of a large and homogeneous population of samples. However, specimens in clinical studies are usually accrued over years and there may be a shift in the participating patient population and also in the methods used to assign disease status owing to the change of practice standards.

The MicroArray Quality Control (MAQC)-II project (Shi et al., 2010) was designed to evaluate these sources of bias in study design by constructing training and validation sets at different times, swapping the test and training sets and also using data from diverse preclinical and clinical scenarios. The goals of MAQC-II were to survey approaches in genomic model development in an attempt to understand the sources of variability in prediction performance and to assess the influences of endpoint signal strength in data. Thus, the use of the MAQC-II datasets can enhance our capability to more accurately predict the clinically relevant cancer prognosis.

The proposed model is evaluated on MAQC-II gene expression data of human breast cancer and multiple myeloma. The breast cancer dataset consisting of 12,993 genes is used to predict pathological complete response (pCR) to preoperative chemotherapy. It was originally divided into two sets: a 130-sample training set consisting of 33 positives and 97 negatives, and a 100-sample test set consisting of 15 positives and 85 negatives. The multiple myeloma dataset consisting of 20,638 genes is used to predict the overall survival (OS) 730 days post-treatment. The original 340-sample training set consisted of 51 positives and 289 negatives, while the 214-sample test set comprised 27 positives and 187 negatives. During preprocessing, sample-wise and feature-wise normalization was conducted, and the variable data values were converted into three-level discretized values {-1, 0, 1} based on z-scores.

The experimental parameter settings are listed in Table 3.1. The parameters are determined as the values yielding optimal performance after empirical experiments. Although a hypergraph classifier has many parameters, most of them can be used as default values independent on problems. Main parameters determined according to problems are initial population size and individual length. Too small initial population causes the discriminative capability of the model to decrease due to the lack of the information for classification. Too large population size leads too heavy computational cost. Therefore, the appropriate range of initial population size is from five to one hundred. Individual length influences the discriminative ability and the probability matching data of a hyperedge. The minimum value of the length is usually set to three and the maximum value does not usually exceed ten. The proper ranges of the parameter values are presented in Table 3.1. To investigate the effect of the Bayesian evolutionary learning method on classification performance, experiments were conducted under various parameter conditions on the model prior.

Table 3.1: Parameter settings of the proposed model used in experiments

Terms	Description	BC dataset	MM dataset
Initial Pop. size	Number of hyperedges	5 **x** $\|D^{tr}\|$	1 **x** $\|D^{tr}\|$
Individual length	Degree of a hyperedge	Min:3, Max:6	Min:3, Max:6
λ	Regularization of the model size	0.001	0.001
ζ	Ratio of MI in fitness value	0.01	0.01
η	Reflecting MI values	1	1
α	Weighting the positive matching function value	0.1	0.1
β	Constant for imbalanced data	1	1
θ_L, θ_C	Matching threshold for learning and classification	0.9, 0.9	0.9, 0.9
R_{max}, R_{min}	Max. and Min. amounts of removed hyperedges	R_{max}: 0.5 **x** $\|E_t\|$ R_{min}: 0.1 **x** $\|E_t\|$	R_{max}: 0.5 **x** $\|E_t\|$ R_{min}: 0.1 **x** $\|E_t\|$
Iteration Number	Condition for terminating the evolution	30	20

BC and MM denote breast cancer and multiple myeloma, respectively.

3.4.2 Prediction Performance

Classification performance was evaluated using six standard classification models: Naive Bayes classifier, random forest (the number of trees = 10), AdaBoost with J48, and support vector machine (SVM) with sequential minimal optimization (SMO) and the second polynomial kernel implemented in Weka (Hall et al., 2009). A variant of learning classifier system (LCS), sUpervised Classifier System (UCS), were also used (Edakunni et al., 2009). We used default values of Weka as the parameters not explained of the other models. In LCS, the pop-size and the iteration number are 1000 and 500, respectively. Because of the large number of variables, probability of the wild card is set to 0.9997. The classification performance of each model was evaluated using the original validation datasets from the MAQC-II project. The results of the evolutionary learning-based models (our model and LCS) were averaged over 10 runs on each test dataset. Prediction performance was based on four measures; sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC), defined below:

$$\begin{aligned}
\text{Sensitivity} &= \tfrac{TP}{TP+FN}, \\
\text{Specificity} &= \tfrac{TN}{FP+TN}, \\
\text{Accuracy} &= \tfrac{TP+TN}{TP+FN+FP+TN}, \\
\text{MCC} &= \tfrac{(TP\cdot TN-FP\cdot FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}
\end{aligned} \tag{3.20}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. In particular, MCC is informative when the ratio of two classes in a dataset is highly skewed. Consequently, MCC has become a popular reference performance measure in bioinformatics, biomedical informatics, and other fields involving unbalanced datasets. MCC values range from +1 to −1, where +1 indicates a perfect prediction, 0 is essentially random prediction, and −1 is the asymptote of extreme misclassification.

Table 3.2 and 3.3 present the performance of the proposed model compared with other models. As revealed by the adjusted p-values, the accuracy of hypergraph

Table 3.2: Comparison of classification performance on breast cancer test dataset

Models	Sensitivity	Specificity	Accuracy	MCC
HC	0.45	0.90	0.84	0.37
2-HC	0 (7.8e-3)	1 (3.9e-3)	0.85 (1.9e-1)	N/A (-)
HC (no MI)	0.04 (7.8e-3)	0.97 (3.9e-3)	0.83 (1.9e-1)	0.11 (7.8e-3)
NB	0.73 (7.8e-3)	0.6 (3.9e-3)	0.62 (9.7e-3)	0.23 (1.7e-2)
RF	0.2 (1.1e-2)	0.81 (3.9e-3)	0.72 (9.7e-3)	0.01 (7.8e-3)
Ada	0.53 (4.5e-1)	0.81 (3.9e-3)	0.77 (9.7e-3)	0.28 (4.9e-2)
SVM	0.46 (4.5e-1)	0.89 (1.5e-1)	0.83 (1.9e-1)	0.35 (6.7e-1)
LCS	0 (7.8e-3)	1 (3.9e-3)	0.85 (1.9e-1)	N/A (-)

HC: Hypergraph Classifiers, 2-HCs: hyperedges of degree 2 (excluding the class vertex), HC (no MI): do not use MI as prior, NB: Naive Bayes, RF: Random Forest, Ada: AdaBoost (J48), SVM: Support Vector Machine, LCS: Learning Classifier System (UCS).

The performance of each model was evaluated using the original test (validation) datasets. All results are averaged over 10 runs on each test dataset. 'N/A' denotes no value and it occurs when TP+FP or FN+TN is zero because a model classifies all data as a certain class label. Values in the parenthesis denote adjusted p-values by multiple comparison correction (Bonferroni correction).

classifiers is similar to those of SVM and LCS, and superior to those of naive Bayes classifier, decision tree, random forest, and AdaBoost on both datasets. The adjusted p-values are calculated based on Wilcoxon signed-ranks test and multiple comparison correction with Bonferroni correction. Compared to existing models, the MCC obtained by our hypergraph model is especially improved on the multiple myeloma dataset with a significant adjusted p-value. Although LCS and SVM

Table 3.3: Comparison of classification performance on multiple myeloma test dataset

Models	Sensitivity	Specificity	Accuracy	MCC
HC	0.33	0.92	0.84	0.32
2-HC	0 (1.9e-3)	1 (5.8e-3)	0.87 (7.8e-3)	N/A (-)
HC (no MI)	0.05 (1.9e-3)	0.97 (5.8e-3)	0.86 (7.8e-3)	0.06 (1.9e-3)
NB	0.55 (1.9e-3)	0.69 (5.8e-3)	0.67 (7.8e-3)	0.17 (1.9e-3)
RF	0.03 (1.9e-3)	0.97 (5.8e-3)	0.85 (5.2e-1)	0.02 (1.9e-3)
Ada	0.22 (1.9e-3)	0.89 (1.4e-1)	0.82 (2.3e-2)	0.22 (1.9e-3)
SVM	0 (1.9e-3)	0.99 (5.8e-3)	0.86 (7.8e-3)	-0.02 (1.9e-3)
LCS	0 (1.9e-3)	1 (5.8e-3)	0.87 (7.8e-3)	N/A (-)

The results are obtained under the same condition as Table 3.2

demonstrate strong prediction accuracy, another measure is necessary for more precisely measuring the prediction capability in these problems, because the accuracy is distorted by severe imbalance of the classes in the datasets. Therefore, the proposed hypergraph classifiers more precisely predict clinical outcomes than existing models in terms of MCC and sensitivity. In addition, comparing the results of HC and 2-HC (a hypergraph classifier with degree-2 hyperedges), we observe that higher-order relationships are more important for accurately predicting cancer prognosis than pair-wise relationships. Moreover, we note that the model performance of HCs using MI as prior is improved by efficient searching of the huge combinatorial space.

Figure 3.4 plots the receiver operating characteristic (ROC) curves of the pro-

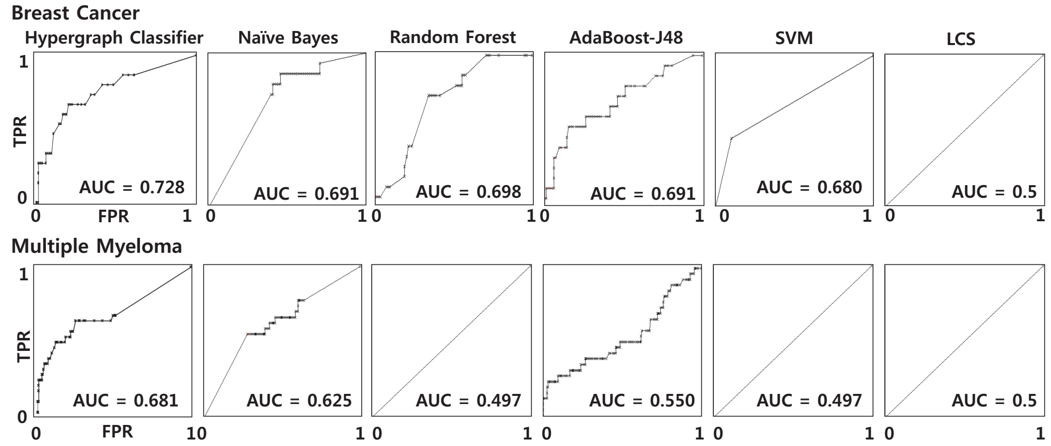


Figure 3.4: ROC curves with AUC of the proposed hypergraph classifier and other models on the test datasets of breast cancer (above) and multiple myeloma (below). TPR (true positive rate) and FPR (false positive rate) denote sensitivity and 1-specificity.

posed hypergraph classifier and other classification models on the test datasets of breast cancer and multiple myeloma, respectively. The areas under the ROC curves (AUCs) are calculated as a measure of predictive discrimination in the given test dataset in terms of specificity and sensitivity. An index of 0.5 presents no discrimination ability, whereas a value of 1 indicates perfect discrimination. Our model showed better classification performance than other models considering AUCs in Figure 3.4 and this result is consistent with that presented in Table 3.2 and 3.3. Interestingly, NB shows relatively high AUC compared to other measures in multiple myeloma and this is caused by the property that an AUC is large when the difference between sensitivity and specificity is small. From these results, we indicate that the hypergraph classifier is suitable model for classifying imbalanced data with

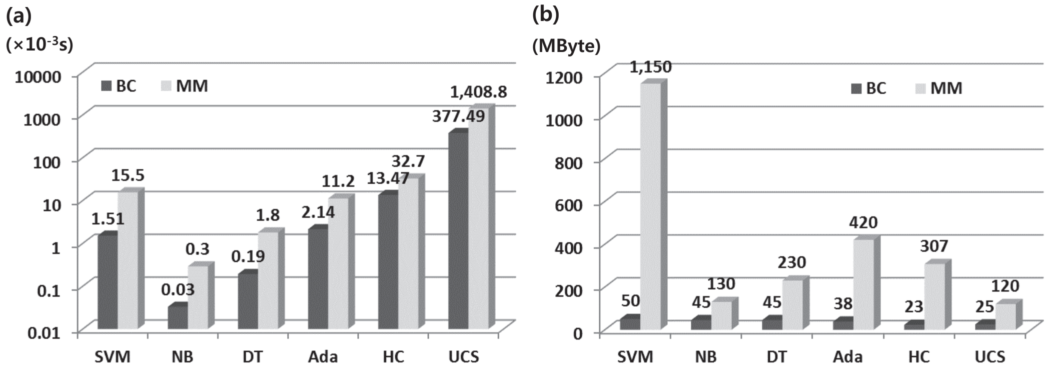
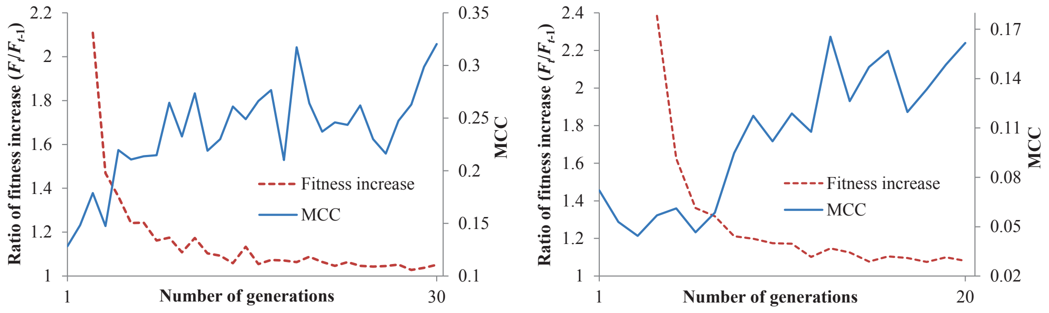


Figure 3.5: The proposed model of (a) time cost and (b) memory size in learning from breast cancer (BC) and multiple myeloma (MM) test data.

high dimensionality compared to other models.

The proposed hypergraph model belongs to a memory-based approach and the model complexity mainly depends on three terms such as the data size, the number of hyperedges, and the hyperedge degrees. Considering that hyperedge degrees can be considered as a constant, the time complexity is O(MN), where M and N denote the number of hyperedges and the data size. Moreover, the number of features increases the model complexity because the size of features usually influences the sufficient number of hyperedges due to the exponential increase of the model space. Figure 3.5 (a) and (b) show the time spent and the memory size used in learning from breast cancer and multiple myeloma dataset, respectively. Our model spends more time in learning compared to other machine learning methods and requires less time than learning classifier systems. The computational environment for the experiments involves Intel Xeon X5690 with 24 cores and 64 Gigabyte RAM based on Window 7 64bit.



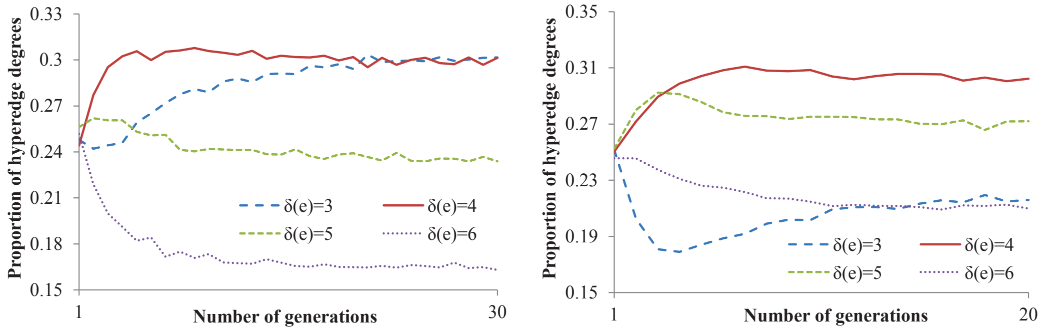
(a) Breast cancer (b) Multiple myeloma

Figure 3.6: MCC fitness dynamics of the evolving hypergraph classifiers, evaluated on test datasets. The results are averages of 10 runs.

3.4.3 Model Analysis

We now present the changes of the proposed model as the Bayesian evolution proceeds. Figure 3.6 shows the dynamics of the MCCs and fitness values evaluated on the breast cancer and multiple myeloma datasets, respectively. Although the MCCs fluctuate, they increase overall as the learning proceeds. The fitness values increase toward their specified maximum. Thus, the defined fitness function reasonably indicates the discriminative capability of the model. In addition, the proposed model evolves into a predictive model that is competitive in terms of both accuracy and MCC despite the skewed class ratio of the data.

Next, we explored the evolution of the hypergraph classifier structure, by analyzing the composition of the hyperedges. The dynamics of hyperedge degree distribution are plotted in Figure 3.7. For both datasets, the proportion of lower-degree hyperedges (δ(e) = 3 and 4) increases as the number of generations increases, while the proportion of higher-order degree hyperedges (δ(e) > 5) decreases. Lower-



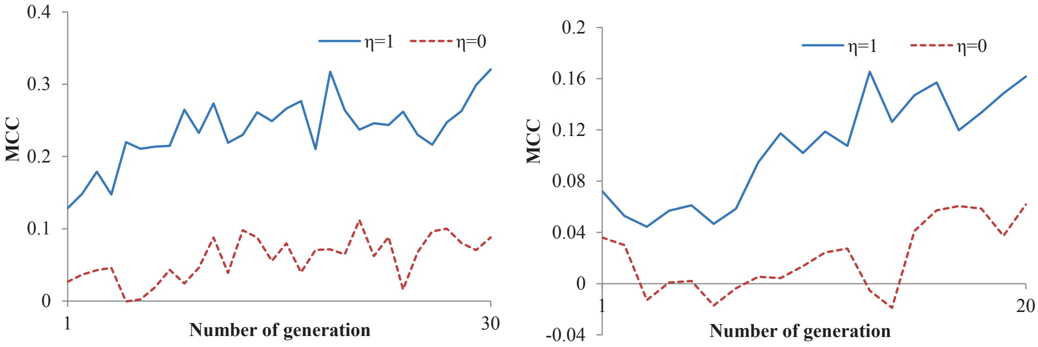
(a) Breast cancer

(b) Multiple myeloma

Figure 3.7: Changes in the distribution of the degree of hyperedges in the evolving hypergraph classifiers. The y-axis denotes the proportion of k-hyperedges to $|E_t|$.

degree hyperedges are assigned a higher weight to reflect their higher probability of matching more training data. In Figure 3.7 (b), especially, 3-hyperedges steadily increase following a decrease in early generations. This initial decrease occurs because, although 3-hyperedges are more likely to match training data, they are also prone to incorrect matching. However, highly discriminative 3-hyperedges survive under the evolutionary learning and thus their proportion increases. Furthermore, higher-degree hyperedges with $\delta(\mathrm{e}) > 5$ are useful for class discrimination because their proportion never converges to zero. Higher-order hyperedges may be especially important for classifying data involving complex relationships between factors. According to Figure 3.7, the proportion of 5-hyperedges ($\delta(\mathrm{e}) = 5$) increases during the early stages of the evolution, and subsequently decreases. This pattern typifies evolutionary phenomena in nature, suggesting that 5-hyperedges play the role of intermediates in the evolutionary process.

Figure 3.8 shows how the learning performance of the model depends on MI



(a) Breast cancer

(b) Multiple myeloma

Figure 3.8: MCC dynamics of the hypergraph classifiers according to MI. $\eta = 0$ denotes that MI as prior was not used.

used as the prior. The effect of the prior on evolving hypergraph classifiers can be investigated by varying the parameter η. From (3.8), when $\eta = 0$, the model reduces to naive random search-based evolution. We observe that MI improves the efficiency of the learning and increases the performance of the model throughout the evolution.

3.4.4 Identification of Prognostic Gene Modules

Here, we analyze the structure of the hypergraph classifiers at the hyperedge level as the model is evolved. Table 3.4 and 3.5 list the genes with large $d(v)$ and the degree of vertices included in hypergraph classifiers learned from each dataset, together with their MI-rank. Genes with large $d(v)$ can be regarded as genes that significantly affect prediction. The threshold of $d(v)$ is defined as the $d(v)$ for which $p < 0.05$, determined by averaging $d(v)$ over all genes. As shown in Table 3.4 and 3.5, many genes with both high and low MI rank appear in the list of large $d(v)$. Those

Table 3.4: List of genes with high $d(v)$ in breast cancer data, identified by the learned model ($p < 0.05$)

♯ App.	Genes (MI-rank)
10	FERMT1(2), SNED1(3), PTGER3(5), HECA(9), MKI67(11), SOX11(12), JMJD6(14), NUCB2(16), FAM153A(19), GREB1(20), TMED7(21), TMEM48(22), KLHDC2(23), GATA3(29), GLI3(31), PIGH(32), CECR5(34), NINJ1(36), DGKG(38), STYXL1(39), DNMT1(43), RASGRP3(44), DEK(45), CLSTN2(46), SCUBE2(50), SLC7A2(52), CSNK1A1(54), SLC16A6(55), VCP(56), MELK(58), TBC1D9(61), KDM4B(67), ASPM(70), ACSM1(76), SKP1(98), ACADVL(78), ADCY1(81), RNF144A(83), BBS4(85),FBXL5(92), UNC119B(95), **TTK(110), AQR(119), MREG(121), VAV3(145), MLPH(164), DNALI1(165), DYRK2(183), YEATS2(200), CCND1(245), PTTG1(252)**
9	MARCH8(1), ASB6(4), GLA(6), CRYZL1(8), IL18R1(24), IRS1(25), CCNE1(27), SOS1(40), CABP2(47), MKL2(51), SMC5(60), ABHD2(65), ORC1(68), JMJD7(86), STK17B(88), PIGH(32), CECR5(34), NINJ1(36), DGKG(38), STYXL1(39), GFRA1(90), **POLDIP3(104), C10orf116(107), BLOC1S1(111), TTC39A(142), PLAGL1(150), TUBGCP4(152), TMSB15B(155), AMFR(163), BLVRA(169), ATPIF1(176), MED13L(192), IGFBP4(198), PJA2(206), MAPT(222), SETD3(229), KIAA0040(243), CENPA(280)**

Appearance number (♯ App.) denotes the number of hypergraph classifiers for which the $d(v)$ of a specific gene is larger than the threshold among the 10 learned models, and thus its maximum value is 10. MI-rank is the rank of the MI value between each gene and the class label. The genes not belonging to top 100 MI rank are bold.

Table 3.5: List of genes with high $d(v)$ in multiple myeloma data, identified by the learned model ($p < 0.05$)

♯ App.	Genes (MI-rank)
10	FSD1(6), HEPACAM(7), CLDN2(13), HSD17B1(53), TDRD3(54), ISOC2(60)
9	LOC100509550(4), ITGAL(17), PREP(19), PGAM2(20), ZMYM1(24), PTDSS2(34), TNNI2(35), QPCT(37), C6orf218(49), SH3KBP1(63), PHLPP1(65), MTMR6(74), FECH(75), RBM45(88), **GGH(102), WHAMM(110),SMAD5OS(125), BPGM(132), NCRNA00208(175), BMP8A(196), GGT7(242), ZACN(258), IFI16(265), CYGB(289), RD3(366), PNKD(375), MOCS3(393), NAT1(581)**

genes with large $d(v)$ but low MI-rank may exert a strong influence on prognostic prediction under the appropriate conditions of other related genes. Moreover, the informative genes repeatedly appear in most of the independently-learned models, indicating that the proposed evolutionary learning method can robustly identify significant hyperedges as prognostic gene modules without the dominant effects of the used prior knowledge. At the same time, the efficiency is enhanced by introducing mutual information to the evolutionary learning of the hypergraph classifiers, without reducing the search space.

Several genes, such as MKI67, CCND1, TTK, PTTG1, CENPA, COX2, and BCL2 have been associated with cancer prognosis in the literature. For example, MKI67 and CCND1 are well-known prognostic markers. They can effectively predict the treatment efficacy of chemotherapy by measuring expression levels of MKI67 and CCND1 (Taneja et al., 2010). TTK and PTTG1 were found to be associated with increased breast cancer risk (Lo et al., 2007). CENPA has also been reported as

Table 3.6: Top 10 gene modules extracted from the learned models in the breast cancer (BC) and multiple myeloma (MM).

BC	
1	[*TTK*^{1}, *ERBB2*^{2}, VAX2]
2	[MFAP1, *CCND1*^{3}, SHCBP1]
3	[GLI3, *PTTG1*^{1}, SOX11, *TTK*^{1}]
4	[C6orf211, NUCB2, *CENPA*^{4}, ZNF207]
5	[ERLIN2, NEK11, *MKI67*^{3}, NAT1]
6	[TTC39A, ABCC4, MFAP5, *MKI67*^{3}]
7	[*CCND1*^{3}, HNRNPM, HOXC6, SNTB1, DGKQ]
8	[CTSL2, *MKI67*^{3}, PDE8B, C16orf42, GLI3]
9	[*MKI67*^{3}, MARCH8, CABP2, SRSF1, BAG1, RTN2]
10	[PSME4, SOS1, DDX58, ELAVL2, SLC16A6, *CENPA*^{4}]
MM	
1	[TEX14, DRAP1, SOX21]
2	[RIOK1, HECW1, CLDN2]
3	[CD58, PAX4, HGFAC, *BCL2*^{5}]
4	[ZNF786, *COX2*^{6}, LOC400128, ANAPC4]
5	[TAX1BP3, *COX1*^{6}, RPL23A, LOC286149]
6	[SFT2D1, FZD5, TMEM11, YTHDF2, *BCL2*^{5}]
7	[EDA, DOC2B, MTMR6, *COX2*^{6}, GMCL1]
8	[MKNK1, UHRF2, MRPL45P2, TMEM160, ATP5J, *BCL2*^{5}]
9	[*COX1*^{6}, POLE2, SPATA18, C14orf153, NSUN6, SLFN5]
10	[CDK17, TMEM42, *COX2*^{6}, LZTS2, RAD51, CARS2]

Genes with a superscript number are confirmed to be related to cancer by the following literature: [1] Lo et al., 2007, [2] Schwaetz et al., 1999, [3] Tanega et al., 2010, [4] McGovern et al., 2013, [5] van de Donk et al., 2006, and [6] Ladetto et al., 2005.

Table 3.7: Gene ontology analysis of the clusters from the learned model in breast cancer (p-value < 0.05)

C	Genes	GO ID	Go Terms	p-value
I	MKI67, MARCH8,	GO:0007049	Cell cycle	7.08e-3
	ACADVL, IL18R1,	GO:0006281	DNA repair	9.63e-3
	TTC39A, SPINLW1,	GO:0048589	Developmental growth	1.29e-2
	BTG2, SMC5, HECA,	GO:0006974	Response to	2.25e-2
	GLI3, BAG4, NEK11,		DNA damage stimulus	
	PSMF1, PDE8B, NOLC1,	GO:0008285	Negative regulation of	2.49e-2
	ERLIN2, PQBP1, NAT1		cell proliferation	
II	CENPA, SCUBE2, ANGEL2,	GO:0007338	Single fertilization	1.00e-2
	ASB4, CUTC, DNALI1,	GO:0006281	Cell motion	4.51e-2
	LHX1 C6orf211, ZNF207	GO:0006974	Cellular developmental process	4.62e-2

a significant independent prognostic marker in patients with ER-positive breast cancer (McGovern et al., 2012). In addition, increased COX2 expression is known as an independent adverse prognostic factor in multiple myeloma (Ladetto et al., 2005). BCL2 is also reported to be associated with the response to interferon therapy in multiple myeloma patients (Donk et al., 2006). Thus, high-degree genes identified by evolutionary learning can be prognostic markers for predicting cancer clinical outcomes, since they form hubs in the learned hypergraph structure. Table 3.6 presents an example of hyperedges as potential gene modules influencing on prognosis prediction. In particular we observe that a module involving TTK and PTTG1 appears concurrently from the learned model in the breast cancer. Interest-

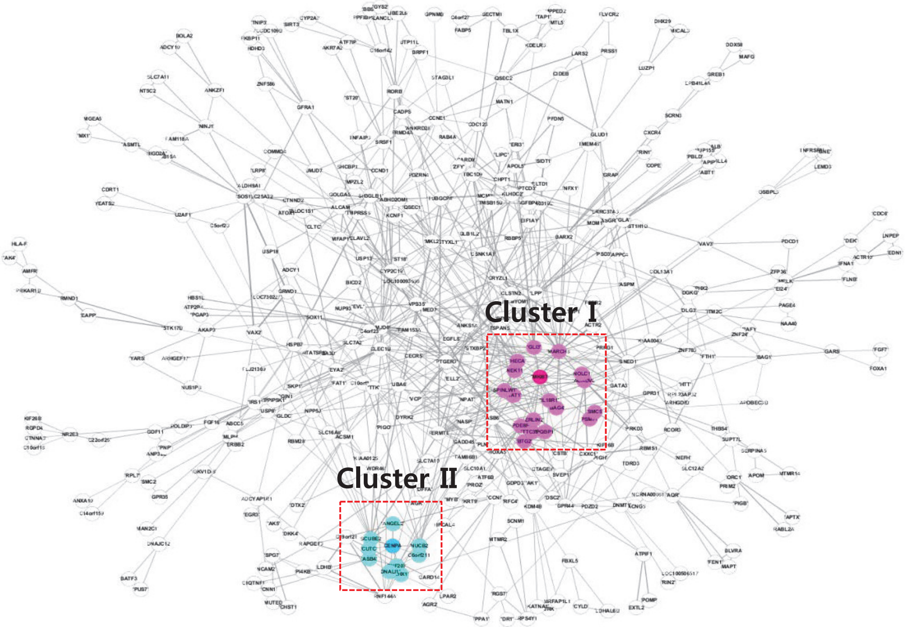


Figure 3.9: Visualization of HCs on breast cancer data. The hypergraph is converted to a normal graph for convenient visualization. This network consists of 422 nodes and 830 edges.

ingly, this finding is consistent with a previous study, in which TTK and PTTG1 act jointly as reproductive risk factors reflecting susceptibility to estrogen exposure for determining breast cancer risk (Lo et al., 2007).

Moreover, the proposed model can be visualized by converting a hyperedge to a clique. Sub-graphs involving genes with large $d(v)$ that are closely related to breast cancer prognosis, such as important prognostic markers, MKI67 and CENPA, are presented in Figure 3.9. In this figure, the cluster is extracted using hypergraph spectral clustering (Zhou et al., 2007), a generalized spectral clustering method (Von, 2007) for hypergraph structures. We also calculated the hypergraph Laplacian L from the learned model, a matrix representing the data variables whose column vectors are eigenvectors of L (Zhou et al., 2007). For clustering, we selected 76 eigenvectors corresponding to eigenvalues below 0.4 from L. Moreover, Table 3.7 shows two gene clusters involved in the network converted from the learned hypergraph with Gene Ontology (GO) analysis (Khatri et al., 2007). The results indicate that genes comprising each cluster have the similar function related to cellular processes. Herein, interpreting the results in this way, we can analyze complex biological phenomena. Thus, the proposed model presents as an alternative method for solving a variety of biomedical problems.

3.5 Summary

We proposed hypergraph classifiers based on evolutionary learning to predict cancer prognoses from complex genetic interactions, using archived data. The learning method evolves a population-based representation of hypergraphs by sequential Bayesian sampling. The Bayesian evolutionary hypergraph model accommodates formal management of model complexity by defining priors on a huge combinatorial search space comprising tens of thousands of genes. Specifically, we controlled the evolutionary search process using two types of prior distributions. One prior guided the compositional variation of the variables in a hyperedge, defined in terms of the mutual information between each genetic variable and the class label. The other was applied on the model size, modulating the degree of a hyperedge and the number of hyperedges in the model.

Cancer prognosis is typically influenced by the combinatorial regulation of multiple genetic factors. By analyzing gene relationships at higher-order levels, we can better predict clinical outcomes in cancer patients. We have demonstrated that higher-order interactions discriminate prognosis more precisely than pair-wise analyzes of single gene relationships. From this viewpoint, we predicted that potential prognostic gene modules could be identified from higher-order gene interactions.

The performance of the proposed method was validated on MAQC-II data. The accuracy of the hypergraph classifiers was similar to that of SVMs and LCSs, and higher than that of naive Bayes classifiers AdaBoost and random forest models. In addition, the MCC of the proposed model was superior to that of existing models. In particular, the MCC score of our model was higher than that of SVMs for multiple myeloma data as 0.34, while the MCC of LCSs was zero for both breast cancer and myeloma datasets. This result indicates that the proposed hypergraph classifiers are robust to imbalanced data, thus more precisely predicting clinical outcomes in cancer patients than existing models. We also compared the performance of the proposed model against two variants of hypergraph classifiers (2-HCs and HCs without using MI as prior). We observe that higher-order relationships are more important for accurately predicting cancer prognosis than pair-wise relationships. Moreover, when hyperedges were generated from information theory, the MCC was improved for both datasets, indicating that searching ability can be enhanced by introducing problem-specific knowledge to the prior in the evolutionary learning process. Furthermore, the interpretable structures of hypergraph classifiers proved useful for analyzing complex biological phenomena. That is, the proposed model presents as an alternative method for solving a variety of biomedical problems. Such contributions will greatly assist toward developing a personalized and refined therapy.

Chapter 4

Hypergraph-based Models for Constructing Higher-Order miRNA-mRNA Interaction Networks in Cancer

4.1 Overview

Dysregulation of genetic factors such as microRNAs (miRNAs) and mRNAs has been widely shown to be associated with cancer progression and development. In particular, miRNAs and mRNAs cooperate to affect biological processes, including tumorigenesis. The complexity of miRNA-mRNA interactions presents a major barrier to identifying their co-regulatory roles and functional effects. Thus, by computationally modeling these complex relationships, it may be possible to infer the gene interaction networks underlying complicated biological processes.

In this chapter, we introduce a data-driven model for identifying cancer stage-

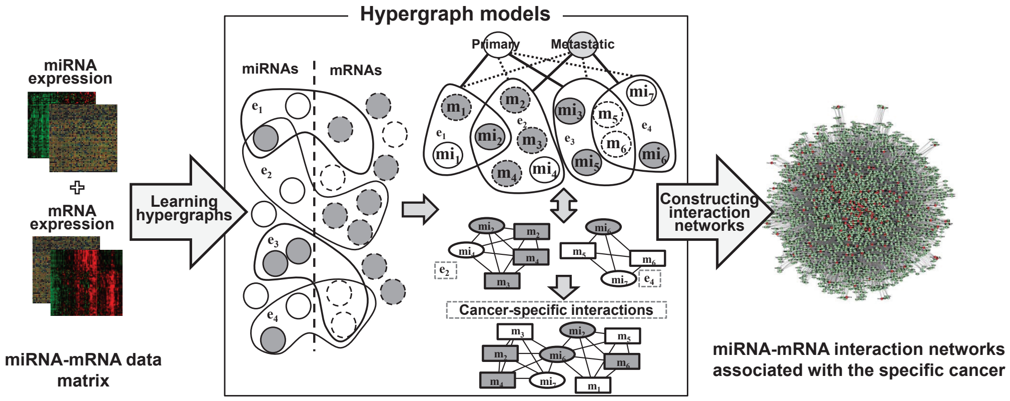

Hypergraph models
miRNA expression
+
mRNA expression
miRNA-mRNA data matrix
Learning hypergraphs
miRNAs
mRNAs
Primary
Metastatic
Cancer-specific interactions
Constructing interaction networks
miRNA-mRNA interaction networks associated with the specific cancer

Figure 4.1: Overview of the hypergraph-based models for constructing higher-order miRNA-mRNA interaction networks at a specific cancer stage. Solid and dotted circles denote miRNAs and mRNAs, respectively. Closed curves denote hyperedges (i.e. modules). In the conventional graph representation (two graphs in the right-bottom of the central box of the figure), ellipses and boxes denote miRNAs and mRNAs, respectively. Grey and white indicate respective high and low gene expression levels.

specific interactions that reflects the high-order relationships between miRNAs and mRNAs (Figure 4.1) (Kim et al., 2012b, 2013b). The proposed model is a hypergraph comprising numerous hyperedges, representing the multi-variable combinations corresponding to miRNAs and mRNAs. Each hyperedge is formally defined as cancer-stage specific statistical figures, and thus our model can deal with real-valued data without discretization. The weight of a hyperedge reflects the strength of the higher-order dependency among the variables of the hyperedge. Therefore, each hyperedge potentially behaves as a gene module. The model explicitly constructs a complex interaction network from many such gene modules. The model is learned by finding a highly-discriminate hypergraph structure from expression profiles using data relevant to a certain stage of prostate cancer.

The learning process involves the iteration of two learning phases; structure and parameter. The structure learning phase constructs a hypergraph of putative hyperedges for discovering potential gene interactions, from a huge feature space represented by the combinations of many miRNAs and mRNAs. Because the miRNA-mRNA interactions are intractably complex, we adopt an evolutionary strategy based on an information theoretic co-regulatory measure, called mutual information. This strategy is used to select genetic variables for generating hyperedges. During the parameter learning phase, the hypergraph is refined by updating the weights of the hyperedges (representing higher-order miRNA-mRNA modules). To this end, we employ a gradient descent method similar to the back-propagation algorithm for learning artificial neural networks. The learned model is then converted into a network structure reflecting the cooperative higher-order gene activities by connecting the extracted hyperedges. Data-driven learning allows the model to build new miRNA-mRNA interaction networks which display the hidden properties of primary and metastatic prostate cancers from a given dataset, which are not known *a priori*.

We construct cancer stage-specific miRNA-mRNA interaction networks reflecting their higher-order relationships using the MSKCC Prostate Oncogenome Project dataset from the model (Taylor et al., 2010). We demonstrate that the proposed model can build several biologically significant miRNA-mRNA interaction networks, including potential modules associated with primary and metastatic prostate cancer. Moreover, cancer-related miRNAs and genes dominate the identified interactions. Some of these interactions, such as hsa-miR-1, hsa-miR-133a, hsa-miR-143, hsa-miR-145, hsa-miR-221, hsa-miR-222, act as hubs in the constructed networks.

We also confirm the biological relevance of the constructed networks through literature review and functional analysis.

4.2 Analyzing Relationships between miRNAs and mRNAs from Heterogeneous Data

Recently, miRNAs have caused great excitement as diagnostic and therapeutic signatures of prostate cancer (Coppola et al., 2010; Gordanpour et al., 2012; Watahiki et al., 2011; Schaefer et al., 2010). They play important roles in cancer pathogenesis, including disease onset, progression, and metastasis, by regulating the stability and translation efficiency of their target mRNAs. Thus, the functional relationships between miRNAs and mRNAs should be elucidated to identify key transcriptional circuits involved in cancer regulation. However, analyzing higher-order miRNA-mRNA relationships is rendered as a challenging problem due to the complexity of their interactions.

Several studies have attempted to identify groups of coherent miRNAs and mRNAs that cooperate in biological processes from heterogeneous data sources via various computational approaches, including probabilistic methods (Yoon and Micheli, 2005; Huang et al., 2006; Joung et al., 2007; Joung and Fei, 2009; Liu et al., 2009a; Bonnet et al., 2010a,b; Liu et al., 2010), rule-based learning (Tran et al., 2008; Liu et al., 2009b), matrix factorization (Zhang et al., 2011), and statistical methods (Peng et al., 2009; Nunez-Iglesias et al., 2010; Lu et al., 2010; Zhang et al., 2012c). These approaches have simplified complex biological mechanisms by systematically analyzing the relationships between genetic elements at the genome level.

Typically, however, bi-relationships between only two factors are assumed in many previous studies (Yoon and Micheli, 2005; Liu et al., 2009b; Zhang et al., 2011;

Peng et al., 2009; Nunez-Iglesias et al., 2010; Lu et al., 2010; Zhang et al., 2012c). Such restrictions are unsuitable for complex genetic interactions because information is lost under the assumption, and biological regulation is controlled by the interaction of multiple genetic components. Many studies have also investigated miRNA-mRNA regulatory interactions using biological information, especially miRNA-target information (Yoon and Micheli, 2005; Huang et al., 2006; Joung et al., 2007; Joung and Fei, 2009; Liu et al., 2009a; Tran et al., 2008; Liu et al., 2009b; Zhang et al., 2011; Peng et al., 2009; Nunez-Iglesias et al., 2010). Biological information reduces the number of false positives, since it provides the predictive model with prior knowledge. In contrast, unknown or hidden interactions not involved in the prior knowledge may be difficult to identify from this information.

To avoid this problem, some probabilistic models which infer miRNA-mRNA modules from expression profiles only, without relying on target information, have been proposed (Bonnet et al., 2010a,b; Liu et al., 2010). Bonnet's model, called LeMoNe (Bonnet et al., 2010a,b) consists of two major steps; the generation of gene clusters based on a feature-sample co-clustering method, and the inference of regulatory modules from generated clusters and regulators based on probabilistically optimized trees. In the clustering approach of Bonnet's method, gene regulatory modules underlying a specific cancer stage are not easily identified. Liu's approach infers functional miRNA regulatory modules using Correspondence Latent Dirichlet Allocation (Corr-LDA) (Liu et al., 2010). The Corr-LDA based model requires discretized data. Since the Corr-LDA model infers probability distributions from latent variables, moreover, miRNAs can be annotated to any functional modules, while mRNAs are restricted to the miRNA-inferred modules.

4.3 Hypergraph-based Models for Identifying miRNA-mRNA Interactions

4.3.1 Hypergraph-based Models

A hypergraph-based model characterizes complex interactions among many genetic factors using hypergraph structures. A hypergraph generalizes the edge concept to a hyperedge by which more than two variables can be connected simultaneously (Zhang, 2008; Kim et al., 2010). As such, it is suitable for representing higher-order relationships among heterogeneous features (e.g. miRNAs and mRNAs). In our model, a hyperedge contains two or more variables corresponding to miRNAs and mRNAs, weighted by the strength of the higher-order dependency among its elements for each class (where the class denotes a specific cancer stage). Thus, each hyperedge implies a set of miRNA-mRNA modules associated with a certain stage of cancer. The proposed model therefore facilitates the construction of higher-order miRNA-mRNA interaction networks among a population of candidate gene modules related to a specific cancer stage.

A hypergraph-based model H is formally defined as a triple $H = (X, Z, E)$ where X, Z, and E denote the sets of miRNAs, mRNAs, and hyperedges, respectively. A hyperedge is represented by a set of statistical values, including mean and covariance for the class label corresponding to a cancer stage. The mean gene expression values differ widely among the class labels, implying that gene expression depends on cancer progression, as shown in Figure 4.2. The hyperedge approach enhances the discriminative capability by combining miRNAs and mRNAs (Figure 4.2). Given an expression dataset with N instances $D = \{d^{(n)}\}_{n=1}^{N} = \{\mathbf{x}^{(n)}, \mathbf{z}^{(n)}, y^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}^{(n)}$ and $\mathbf{z}^{(n)}$ are real-valued vectors of miRNA and mRNA expressions in the n-th instance, and y is an element of a cancer stage set Y, the i-th hyperedge e_i

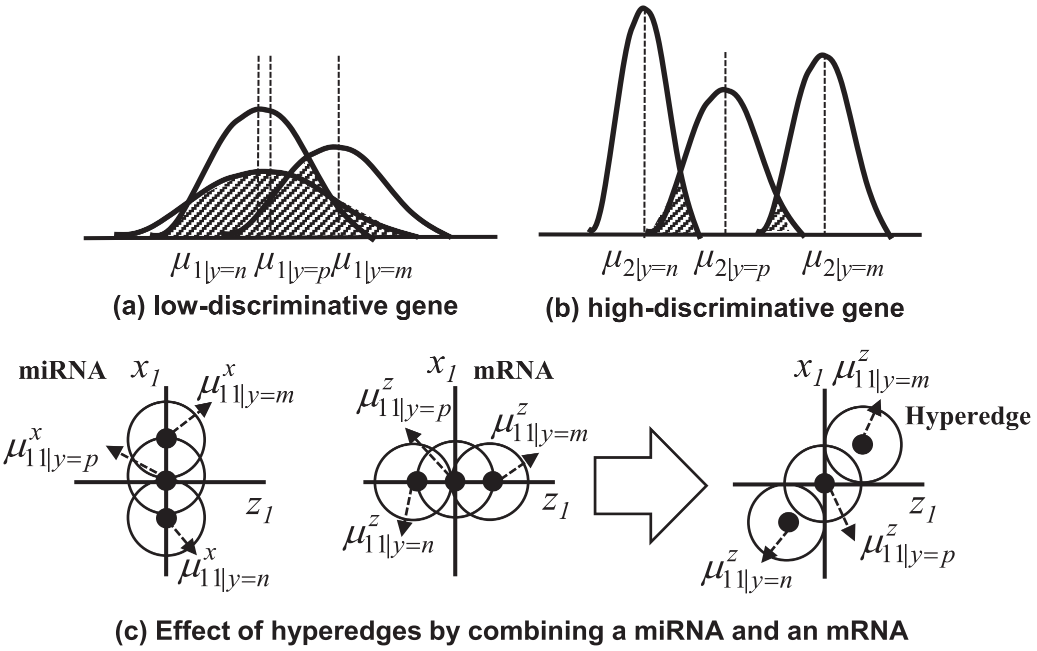


Figure 4.2: Biological meaning of mean and variance used in representing a hyperedge. Panels (a) and (b) illustrate how the means and variances differ between low and high discriminative genetic factors. A gene is low-discriminative when the means are similar at each disease stage but the variances are large (where n, p, and m denote normal, primary, and metastatic stage, respectively). Panel (c) illustrates the enhanced discriminative capability of a hyperedge involving two genetic factors. By comparing the discriminative capability of each miRNA or mRNA, the discrimination capability of the hyperedge is enhanced.

contains the mean vectors and the covariance of its miRNAs and mRNAs for the given cancer stage:

$$e_i = \left\{ \begin{array}{c} e_{i|y=y_1} \\ ... \\ e_{i|y=y_{|Y|}} \end{array} \right\} = \left\{ \begin{array}{c} (\mu_i, \Sigma_i)_{|y=y_1} \\ ... \\ (\mu_i, \Sigma_i)_{|y=y_{|Y|}} \end{array} \right\}, \tag{4.1}$$

$$\mu_i = (\mu^x_{i1}, ..., \mu^x_{il}, \mu^z_{i1}, ..., \mu^z_{im}) \text{ and } l + m = |e_i| \tag{4.2}$$

where μ^x_{ij} and μ^z_{ik} denote the means calculated from the expression profiles of the j-th miRNA and the k-th mRNA, respectively, in the i-th hyperedge (whose elements comprise l miRNA and m mRNAs). l and m are called the degrees of miRNA and mRNA of the hyperedge, respectively. By the definition of a hyperedge, each hyperedge has $|Y|$ mean vector/covariance pairs, and $|Y|$ weights. The hypergraph-based model is considered as a population of hyperedges. Given a gene expression profile ($\mathbf{x}$, $\mathbf{z}$), the cancer stage of the profile is classified as y^*, for which the summation of the expected values (the products of the hyperedge weight and the probability of ($\mathbf{x}$, $\mathbf{z}$) matching the hyperedge), is highest among the elements of Y. "($\mathbf{x}, \mathbf{z}$) matches $e_i|y$" means that ($\mathbf{x}$, $\mathbf{z}$) has similar expression values to ones of the i-th hyperedge with respect to the genetic variables involved in $e_i|y$ at cancer stage y, and we introduce a Gaussian kernel into the hyperedge to calculate the matching probability of ($\mathbf{x}$, $\mathbf{z}$) and $e_{i|y}$, $P(u = 1|\mathbf{x}, \mathbf{z}, e_{i|y})$. The matching probability is calculated by the normalized subdimensional distance between $e_{i|y}$ and ($\mathbf{x}$, $\mathbf{z}$):

$$P(u = 1|x, z, e_{i|y}) = \exp\left\{-\beta d(x, z, e_{i|y})\right\}, \tag{4.3}$$

$$d(x,z,e_{i|y}) = \frac{1}{|e_i|}\left\{\sum_{j=1}^{l}\frac{\left(x_{ij}-\mu_{ij}^{x}\right)^2}{\left(\sigma_{ij|y}^{x}\right)^2}+\sum_{k=1}^{m}\frac{\left(z_{ik}-\mu_{ik}^{z}\right)^2}{\left(\sigma_{ik|y}^{z}\right)^2}\right\}^{\frac{1}{2}}, \tag{4.4}$$

where u=1 denotes that (**x**, **z**) matches $e_{i|y}$, $\sigma^x_{ij|y}$ and $\sigma^z_{ij|y}$ are the standard deviations of x_{ij} and z_{ik} (the j-th miRNA and k-th mRNA, respectively) in the i-th hyperedge for a given y, and β is a constant for adjusting the probability. Larger β implies smaller matching probability, and therefore a smaller number of hyperedges influence on classifying the data. Specifically, the cancer stage y^* of (**x**, **z**) is computed as follows:

1. Calculate $c_{y'}$, the sum of the expected values for each y' in Y over all hyperedges of H:

   $$c_{y'} = \sum_{i=1}^{|H|} w(e_{i|y=y'})P(u = 1|x, z, e_{i|y=y'}) \tag{4.5}$$

   where $|H|$ denotes the number of hyperedges and $w(e_{i|y})$ is the weight of $e_{i|y}$, explained in the next subsection.

2. Predict the cancer stage as y^*:

   $$y^* = \arg\max_{y' \in Y} c_{y'} \,. \tag{4.6}$$

In terms of distance-based connectionist models, our model is related to radial basis function networks (RBFNs) (Buhmann, 2003). Whereas RBFNs use kernelized distance for all variables, the proposed hypergraph model uses the probability derived from the subdimensional distance on the projected space corresponding to each hyperedge. Unlike RBFNs, therefore, the hypergraph model can detect embedded subpatterns reflecting higher-order relationships among the components. Because these embedded subpatterns influence the classification, we can intuitively analyze the complex interactions of genetic factors that contribute to classifying a specific cancer stage.

4.3.2 Learning Hypergraph-based Models

The proposed model learns by finding a hypergraph structure with high discriminative capability at a specific cancer stage. This is achieved by maximizing the conditional likelihood for a model H and the gene expression profiles and a log function is adopted for convenience. To minimize the error of classifying the cancer stage, $E_{D,H}$, the log conditional likelihood is maximized by least mean square criteria using (4.5) and a sigmoidal function:

$$\begin{aligned}
&H* = \arg\max_H \log \prod_{n=1}^{N} p(y^{(n)}|x^{(n)}, z^{(n)}, H) = \arg\max_H \sum_{n=1}^{N} \log p(y^{(n)}|x^{(n)}, z^{(n)}, H) \\
&\equiv \arg\max_H \sum_{n=1}^{N} \delta(y^{(n)}, y'_H) = \arg\min_H E_{D,H} \qquad (4.7) \\
&\approx \arg\min_H \sum_{n=1}^{N} \sum_{y' \in Y} \left(\delta(y^{(n)}, y') - P(y'|x^{(n)}, z^{(n)}, H)\right)^2
\end{aligned}$$

s.t.

$$P(y'|x,z,H) = \left\{1 + \exp\left(c_{y'} - \frac{1}{|Y|} \cdot \sum_{y \in Y} c_y\right)\right\}^{-1}$$

where ($\mathbf{x}^{(n)}$, $\mathbf{z}^{(n)}$) denotes the n-th miRNA-mRNA expression and $y^{(n)}$ is the cancer stage of the example. y'_H is the label predicted by H and $\delta(y^{(n)}, y'_H)$ is an indicator function, equal to 1 if $y^{(n)}$ equals y'_H, and 0 otherwise. To enhance the classification accuracy, it is essential that the population comprises hyperedges with high discriminative capability, and the hyperedge weights must be refined to minimize (4.7) in the generated hypergraph.

To meet these requirements, the learning iterates two phases: structure learning and parameter learning. The structure learning constructs a hypergraph from hyperedges that identify potential miRNA-mRNA modules. The weights of the hyperedges are updated to minimize the classification error of the generated gene module population during the parameter learning phase. Because the hypergraph model represents a huge combinatorial feature space (size $2^{|x|+|z|}$) of many miRNAs and mRNAs, exhaustively searching for the optimal population is infeasible. Instead we adopt an evolutionary learning method based on information-theoretic criteria to generate putative hyperedges for the structure learning.

We assume that a hyperedge consisting of strongly interactive miRNAs and mRNAs is highly discriminative for classification in this study. Mutual information is used as a co-regulatory measuring criterion for efficiently selecting genes for hyperedge generation. Mutual information (MI) is an information-theoretic measure that specifies the degree of conditional independency between two random variables. When a genetic factor more strongly determines the cancer stage, the MI between the gene and the cancer stage is increased. A hyperedge is generated by probabilistically selecting miRNAs and mRNAs, and the MI between each gene and the class label determines the probability of selecting the genes. The probability $P_I(X_i)$ of selecting the i-th gene X_i is defined such that miRNAs or mRNAs with high MI are selected more frequently:

$$P_I(X_i) = \frac{\{I(X_i;Y)\}^{\eta}}{\sum\limits_{X_i \in X} \{I(X_i;Y)\}^{\eta}}, \tag{4.8}$$

where denotes the MI between the i-th genetic factor and the cancer stage, and η is a nonnegative constant that regularizes the influence of MIs on the gene selection. When η is zero, all variables may be selected with equal probability. Once the hyperedges have been generated, the mean vectors and covariance of the hyperedges are calculated from the training dataset. To identify putative strongly-interacting miRNA-mRNA modules, the initial weight of the *i*-th hyperedge is computed using the variances of each genetic factor and the multivariate MI (Kraskov et al., 2004) among all variables, including the class label involved in the hyperedge. A gene with a particular mean expression value but small variance likely possesses higher discriminative capability than one with larger variance. Moreover, by the definition of MI, large multivariate MI implies more relationships among the genes. Thus the initial weight of a hyperedge is defined as

$$w_0(e_{i|y}) = \kappa \cdot I(e_i) + \sum_{x_{ij} \in e_i} \frac{1}{\sigma^2_{ij_{|y}}} \tag{4.9}$$

s.t.

$$\begin{aligned} I(e_i) &= I(X_{i1,};..;X_{ik};Y) = I(X_{i1,};..;X_{ik}) - I(X_{i1,};..;X_{ik}|Y) \\ &= I(X_{i1,};..;X_{ik}) - E_Y(I(X_{i1,};..;X_{ik})|Y)\ , \end{aligned}$$

where k is the number of variables of e_i and κ denotes the ratio of the variance to MI. In the parameter learning phase, the weights of the hyperedges are updated using the gradient descent method for all training data. The aim is to minimize the error in terms of the classification probability in (4.3) and the matching probability in (4.7):

$$w_t(e_{i|y}) = \Delta w_{t,i|y} + w_{t-1}(e_{i|y}), \tag{4.10}$$

$$\Delta w_{t,i|y} = \frac{\gamma}{t} P(y|\mathbf{x},\mathbf{z},H)\,(1 - P(y|\mathbf{x},\mathbf{z},H))\,(\delta(\tilde{y},y) - P(y|\mathbf{x},\mathbf{z},H)) \cdot P(u = 1|\mathbf{x},\mathbf{z},e_{i|y}),$$

where $\tilde{y}$ is the real cancer stage of a miRNA-mRNA expression sample, and t and γ denote the epoch number in the parameter learning and the parameter learning rate, respectively. The epoch is the number of weight updates for the built hypergraph during parameter learning, and γ controls the extent of weight change during parameter learning. Thus, the weight becomes high when the hyperedge consists of miRNAs and mRNAs with strong higher-order interactions and when the variances of the gene variables are small at all cancer stages. Following parameter learning, low weighted hyperedges are removed from the population, and the next structure learning step is performed. To prevent the removal of highly discriminating hyperedges, the number of replaced hyperedges decreases to a specific value as the iterations proceed, as follows:

$$R_t = \frac{R_{max} - R_{min}}{\exp(t)} + R_{min}, \tag{4.11}$$

where t is the iteration number of the structure learning phase, and R_{max} and R_{min} denote the maximum and minimum number of replaced hyperedges, respectively. Therefore, the number of replaced hyperedges consecutively decreases as the structure learning proceeds, while high-discriminative modules are preserved. The algorithm for learning the hypergraph-based model is presented in Figure 4.3.

4.3.3 Building Interaction Networks from Hypergraphs

We construct a higher-order miRNA-mRNA interaction network at a specific cancer stage from the learned model. When analyzing complex biological networks based on graph mining, frequently occurring subgraphs in the networks are generally regarded as important building blocks which are merged to create the functional network (Hu et al., 2005; Mason and Verwoerd, 2007; Yan et al., 2007; Ramadan et al., 2010). Since a high-weight hyperedge corresponds to a significant subgraph

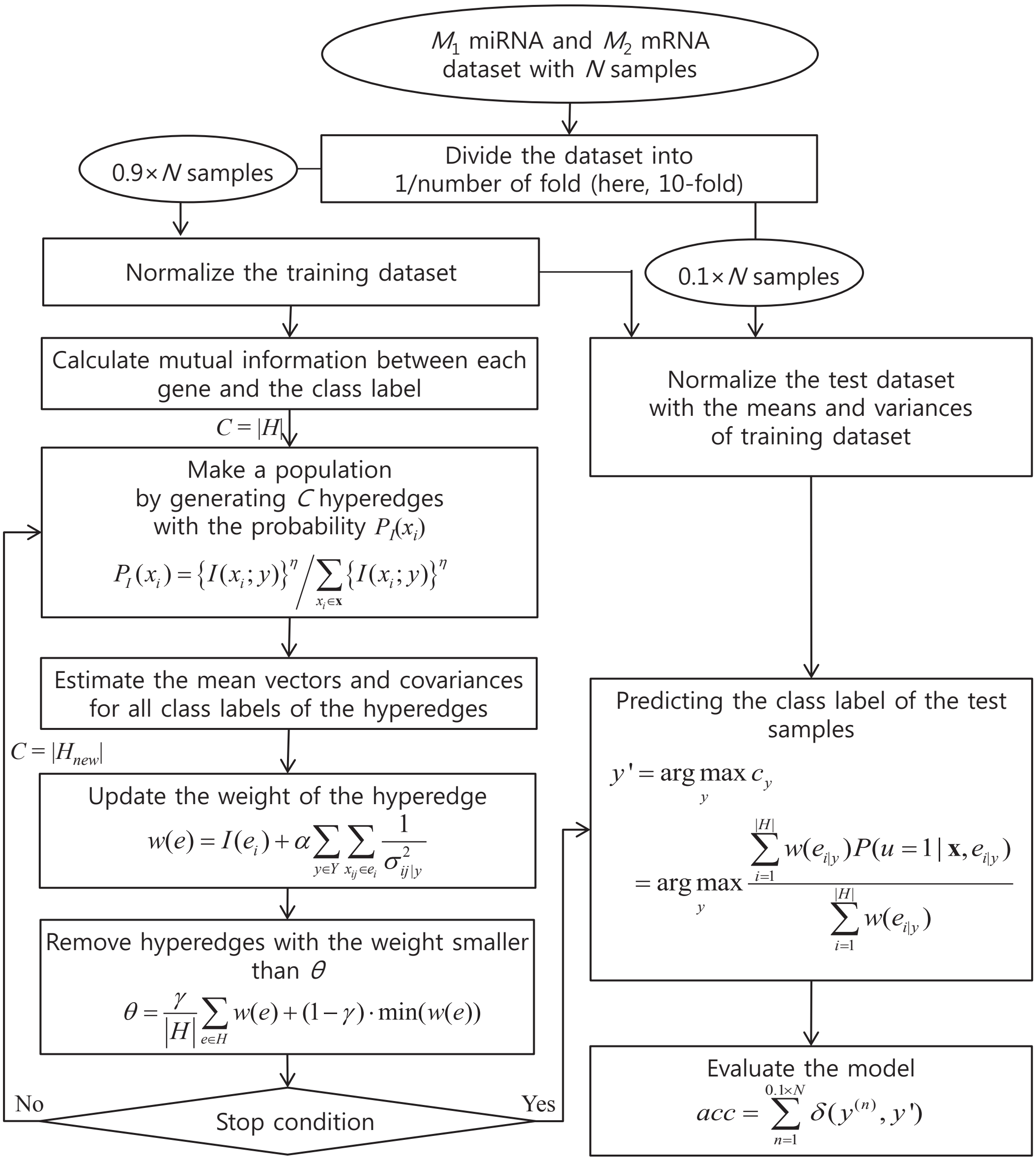

M_1 miRNA and M_2 mRNA dataset with N samples
Divide the dataset into 1/number of fold (here, 10-fold)
$0.9\times N$ samples
Normalize the training dataset
$0.1\times N$ samples
Calculate mutual information between each gene and the class label
Normalize the test dataset with the means and variances of training dataset
$C = |H|$
Make a population by generating C hyperedges with the probability $P_I(x_i)$
$P_I(x_i) = \{I(x_i;y)\}^\eta \Big/ \sum_{x_i\in\mathbf{x}} \{I(x_i;y)\}^\eta$
Estimate the mean vectors and covariances for all class labels of the hyperedges
$C = |H_{new}|$
Update the weight of the hyperedge
$w(e) = I(e_i) + \alpha \sum_{y\in Y}\sum_{x_{ij}\in e_i} \frac{1}{\sigma^2_{ij|y}}$
Predicting the class label of the test samples
$y' = \arg\max_y c_y$
$= \arg\max_y \frac{\sum_{i=1}^{|H|} w(e_{i|y}) P(u=1\,|\,\mathbf{x}, e_{i|y})}{\sum_{i=1}^{|H|} w(e_{i|y})}$
Remove hyperedges with the weight smaller than θ
$\theta = \frac{\gamma}{|H|}\sum_{e\in H} w(e) + (1-\gamma)\cdot\min(w(e))$
No
Stop condition
Yes
Evaluate the model
$acc = \sum_{n=1}^{0.1\times N} \delta(y^{(n)}, y')$

Figure 4.3: Algorithm for learning the hypergraph-based model

reflecting a higher-order relationship among genetic variables, the interaction network is constructed by connecting cliques sharing common genes. A hyperedge is assigned separate weights for each cancer stage and it is merged into the graph of the highest weighted cancer stage. Formally, a cancer-stage and a cancer stage-specific interaction network $G_{|y'}$ =(V, E), where V and E denote a vertex set and an edge set, respectively, is constructed by merging the hyperedges as follows (where y' is the class label with the largest weight value):

$$G_{|y'} = G_{|y'} \cup C_i, \tag{4.12}$$

$$y' = \arg\max_{y \in Y} \left\{ w(e_{i|y}) \right\}, \tag{4.13}$$

and C_i is a clique corresponding to the i-th hyperedge e_i (Figure 4.4). This dividing and remerging approach enables the constructed interaction networks to be easy-to-visualized without impairing the higher-order property of the model since the weight of edges in the constructed networks are derived from the hyperedge weights reflecting the strength of the higher-order interaction.

4.4 Constructing miRNA-mRNA Interaction Networks Based on Higher-Order Relationships

4.4.1 Data and Experimental Settings

The clinical heterogeneity of prostate cancer, coupled with its high prevalence, raises challenges in the management of newly diagnosed patients as well as those with metastatic disease. Specifically, prostate cancer shows enormous biological

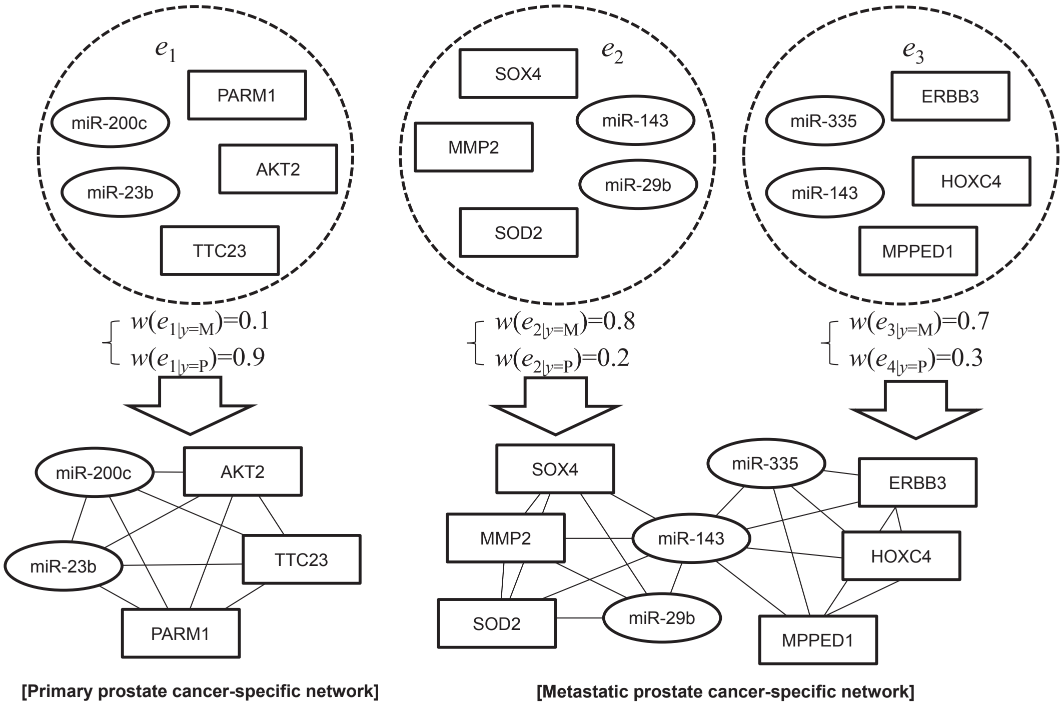

e_1
PARM1
miR-200c
AKT2
miR-23b
TTC23
e_2
SOX4
miR-143
MMP2
miR-29b
SOD2
e_3
ERBB3
miR-335
HOXC4
miR-143
MPPED1
$w(e_{1|y=\mathrm{M}})$=0.1
$w(e_{1|y=\mathrm{P}})$=0.9
$w(e_{2|y=\mathrm{M}})$=0.8
$w(e_{2|y=\mathrm{P}})$=0.2
$w(e_{3|y=\mathrm{M}})$=0.7
$w(e_{4|y=\mathrm{P}})$=0.3
miR-200c
AKT2
miR-23b
TTC23
PARM1
SOX4
MMP2
miR-143
SOD2
miR-29b
miR-335
ERBB3
HOXC4
MPPED1
[Primary prostate cancer-specific network]
[Metastatic prostate cancer-specific network]

Figure 4.4: Procedure of converting a hypergraph to cancer stage-specific interaction networks. ′P′ and ′M′ denote metastatic and primary prostate cancer, respectively.

heterogeneity, with some patients dying of metastatic disease within 2-3 years of diagnosis whereas others can live for 10-20 years with organ-confined disease, likely a reflection of underlying genomic diversity. Herein, understanding prostate cancer mechanisms requires integrated large-scale cancer genomic projects which can provide new insights into the molecular classification of cancers. In particular, miRNAs have been recognized as the key regulator of gene expression in prostate cancer. Thus, the integrated analysis of miRNA and mRNA expression on a genome-wide level can offer more informed clinical decision-making and novel therapeutic

Table 4.1: Parameter settings for experiments

Parameters	Values	Parameters	Values
♯ of miRNAs	3	♯ of mRNAs	5
♯ of modules	variable	β in (4.3)	1.0
Epochs of structure learning	100	Epochs of parameter learning	20
η in (10)	1.0	κ in (4.9)	1.0
γ in (13)	1.0	R_{max}, R_{min}	0.9, 0.5

targets.

In this study, miRNA and mRNA expression profiles obtained from the MSKCC Prostate Oncogenome Project (Taylor et al., 2010) were matched at three stages of prostate cancer. The dataset contains 373 miRNAs and 19,780 mRNAs from 27 normal, 98 primary and 13 metastatic stages. During preprocessing, sample-wise and feature-wise normalization was conducted, and miRNAs and mRNAs were separately normalized. The experimental parameter settings are listed in Table 4.1. The parameters are those yielding optimal performance in empirical experiments. A hypergraph can include hyperedges with different number of genetic variables but we fixed the number of variables for all hyperedges of a hypergraph in this study.

4.4.2 Classification Performance

Classification performance was evaluated using three standard classification models; support vector machines (SVMs) with the 2nd polynomial kernel and sequential minimal optimization (SMO), *k*-th nearest neighbor classifiers (*k*-NNs), and naive

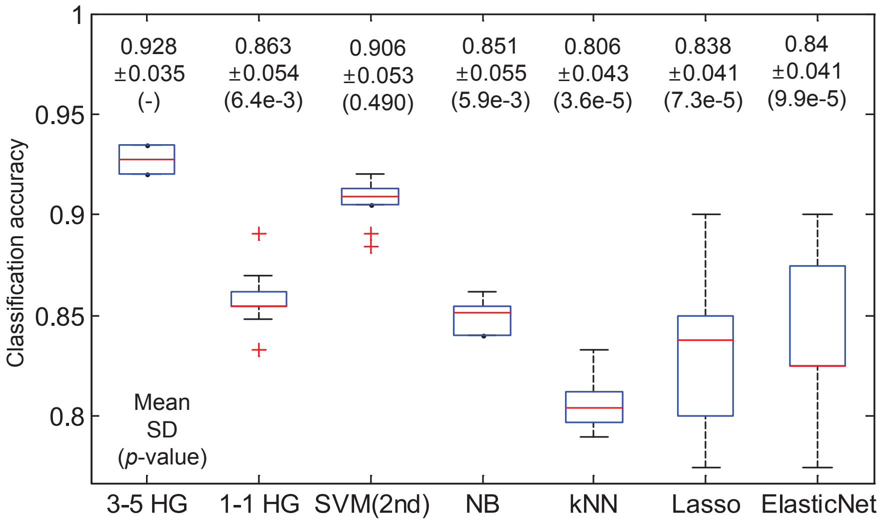


Figure 4.5: Boxplots of classification accuracy on the test set. it m-*n* HG denotes the hypergraph-based model whose all hyperedges embody *m* miRNAs and *n* mRNAs. All results are averaged after 10 runs by 10-fold cross validation. *P*-values are calculated using *t*-test of our model and other models.

Bayes classifiers (NBs) implemented in Weka (Hall et al., 2009). The MATLAB algorithms lasso and elastic net (α=0.5) were also used. All results were averaged over 10 experiments. Figure 4.5 presents the classification accuracy of our model compared to other models. As revealed by the *p*-values of the *t*-test, the proposed hypergraph-based model competes on-par with SVMs and outperforms the *k*-NN, NB and Lasso-based methods. In addition, by comparing the results of 3-5 HG (a hypergraph model whose hyperedges consist of three miRNAs and five mRNAs) and 1-1 HG, we observe that higher-order relationships are more important for discriminating cancer stages than pair-wise relationships between a single miRNA and mRNA.

4.4.3 Model Evaluation

The proposed hypergraph-based learning method is evaluated on simulation data for verifying whether the method finds true solutions. The data consist of 500 instances with 7 variables whose mean is zero and the class label of each instance is determined as follows:

$$x_i \sim N(0,1),\ \ 1 \leq i \leq 7$$

$$c^{(n)} = \begin{cases} 1, & if\, x_2 > 2\ \wedge\ x_3 > 2\ \wedge\ x_4 > 2 \\ 2, & if\, x_5 < -2\ \wedge\ x_6 < -2\ \wedge\ x_7 < -2 \\ \quad 3, & otherwise \end{cases}, \tag{4.14}$$

where x_i and $c^{(n)}$ denote the i-th random variable and the class label of the n-th instance. Table 4.2 illustrates the classification accuracy and predefined modules in the learned model. The accuracy is averaged after 10 experiments by 10-fold cross

Table 4.2: Verification result on the simulation dataset

Models	SVM	DT	kNN	HG	Module 1	Module 2
Accuracy	0.956	0.886	0.93	0.956	10	10
±SD	±0.002	±0.004	±0.006	±0.003	-	-

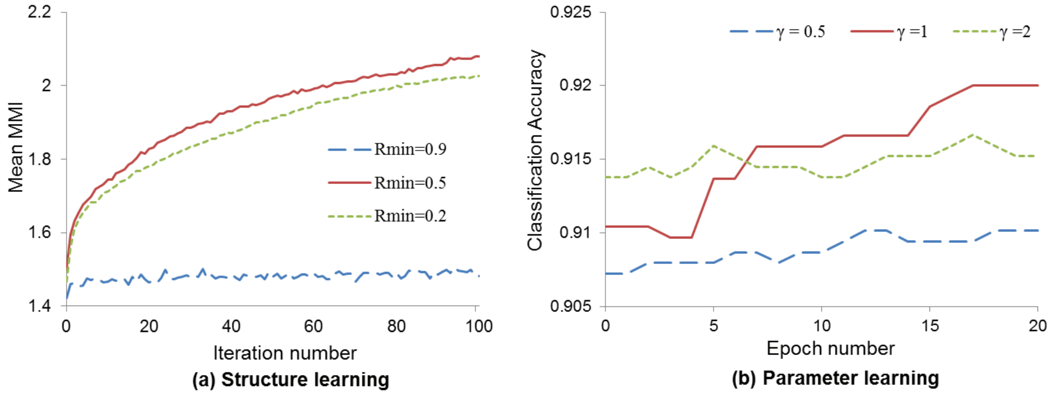


Figure 4.6: Learning curves in the structure and the parameter learning phases. As the performance measure, we used mean multivariate mutual information (MMI) of all hyperedges in the model for the structure learning and accuracy on 10-fold cross validation for the parameter learning. *Rmax* is fixed as 0.9 in (a) and γ is a learning rate for the parameter learning in (b). All results are averaged on 10 experiments of 10- fold cross validation.

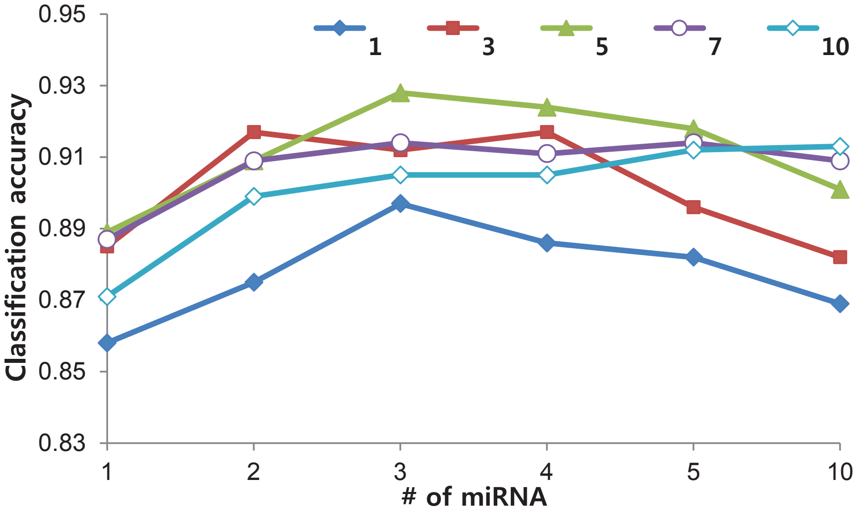
validation, and each hypergraph includes 20 hyperedges with four variables. In Table 2, Module 1 and 2 means the number of case when there exist hyperedges involving a predefined-set 1 (x_2, x_3, x_4) and 2 (x_5, x_6, x_7) in a learned hypergraph. Because we conducted 10-fold cross validation, the maximum values of Module 1 and 2 are ten. Therefore, we indicate that our method can find true solutions from small combinatorial spaces, considering the accuracy and the number of found variable modules.

Figure 4.6 presents two learning curves under various conditions of the structure (a) and the parameter (b) learning phases. As the measure for structure learning, we used mean multivariate mutual information (MMI) of all hyperedges in the model because the goal of the structure learning is to find the significant higher-order cancer-specific gene interaction modules, and an MMI is the measure reflecting the strength of interactions among genetic factors in the hyperedges considering the stage of cancer. On the other hand, classification accuracy is used as the measure for the parameter learning phase since the weight for each cancer stage is updated to minimize the error in the phase. Figure 4.6 (a) presents the increase of mean MMI under various *Rmin* which is the minimum ratio of the hyperedges replaced in the iteration, and plays a role of the structure learning rate. We indicate that too large an *Rmin* causes low MMI by replacing too many hyperedges and too small an *Rmin* leads slow increase of the MMI from Figure 4.6 (a). Figure 4.6 (b) presents similar results to (a) with respect to the effect of learning rate γ.

Moreover, Figure 4.7 shows the classification accuracy according to the number of genetic factors in the hyperedges. The classification accuracy is the best when a hypergraph consists of hyperedges with three miRNAs and five mRNAs. We indicate that small number of genetic variables show worse performance because various processes of prostate cancer is influenced on the complex interactions among many features. Furthermore, the accuracy of the hypergraphs including hyperedges with more than ten genetic variables is low since the models consist of too specific information and thus have the low generalization property.

Figure 4.8 shows that the proposed learning method can stably extract significant genetic factors despite its random selection approach. We define a measure as the number of appearance of a gene in the model, $A(x_i)$, for verifying the stability of the model as follows:

$$A(x_i) = \sum_{m=1}^{100} \delta(x_i, H_m)$$



mRNA \ miRNA	1	2	3	4	5	10
1	0.858(0.015)	0.875(0.009)	0.897(0.012)	0.886(0.019)	0.882(0.01)	0.869(0.01)
3	0.885(0.015)	0.917(0.011)	0.912(0.01)	0.917(0.005)	0.896(0.01)	0.882(0.011)
5	0.889(0.011)	0.909(0.011)	**0.928(0.007)**	0.924(0.006)	0.918(0.009)	0.901(0.008)
7	0.887(0.01)	0.909(0.005)	0.914(0.006)	0.911(0.005)	0.914(0.008)	0.909(0.008)
10	0.871(0.007)	0.899(0.007)	0.905(0.011)	0.905(0.007)	0.912(0.008)	0.913(0.011)

accuracy (±SD)

Figure 4.7: Classification accuracy according to the number of miRNA and mRNA in the hyperedges. The classification accuracy is the best when a hypergraph consists of hyperedges with three miRNAs and five mRNAs. All results are averaged on 10 experiments of 10-fold cross validation.

$$\delta(x_i, H_m) = \begin{cases} 0 \mathrm{if} x_i \mathrm{isnotinvolvedin} H_m \\ 1 \ \ otherwise \end{cases}, \tag{4.15}$$

where x_i denotes the i-th miRNA or mRNA, and H_m is the m-th learned model. $\delta(x_i, H_m)$ is an indicator function and it returns one when xi appears at least once in H_m, otherwise zero. The proposed method is compared to randomly generated hypergraphs each comprising 200 hyperedges involving three miRNAs and five mRNAs. The results are derived from 100 models learned by 10 experiments of 10-fold cross validations, and 100 randomly generated hypergraphs. According to Figure 4.8 (a), our method extracts significant miRNAs only, while almost all of the miRNAs are involved in random graphs. Moreover, whereas the learning method selects several significant mRNAs, all mRNAs appear at low frequency in the random graphs, as shown to Figure 4.8 (b). The stability and reproducibility of the proposed model is evident from the high-frequency occurrence of high ranked miRNAs and mRNAs, indicating that certain genes persist in the models.

4.4.4 Constructed Higher-Order miRNA-mRNA Interaction Networks in Prostate Cancer

The miRNA-mRNA interaction network constructed from the proposed model is illustrated in Figure 4.9 and 4.10 for primary and metastatic prostate cancer respectively (Smoot et al., 2011). The constructed interaction networks comprise putative miRNA-mRNA modules associated with each stage of prostate cancer, and reflect their higher-order relationships. The primary prostate cancer network includes 67 miRNAs and 233 mRNAs, while the metastatic prostate cancer network involves 65 miRNAs and 180 mRNAs.

Many of the miRNAs in the constructed networks have been significantly asso-

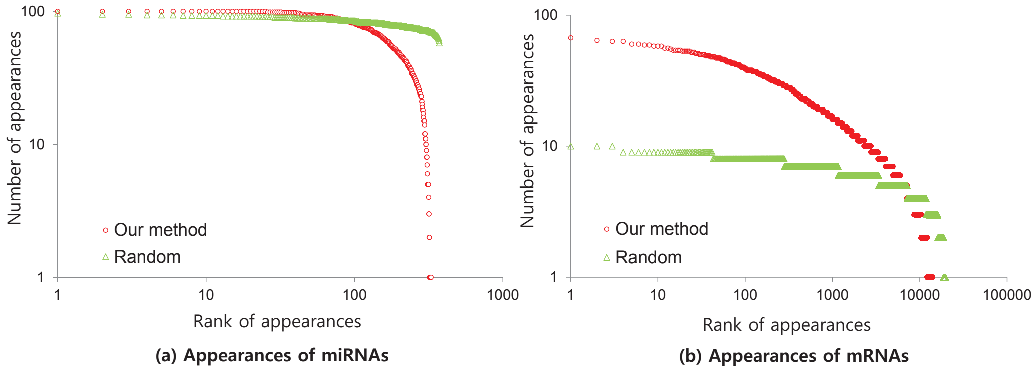


Figure 4.8: Reproducibility of decisive miRNAs (a) and mRNAs (b) influencing on classification. 100 hypergraphs are generated by randomly selecting miRNAs and genes, while another 100 hypergraphs are generated by our learning method (10 experiments with 10-fold cross validation). Each hypergraph includes 200 hyperedges consisting of three miRNAs and five mRNAs. The x-axis denotes the rank of the appearance of miRNAs or mRNAs, and y-axis is the number of miRNA or mRNA appearances. Both axes are log-scaled.

ciated with prostate cancer in the literature, and are thus termed prostate cancer-related miRNAs (Jiang et al., 2009). In addition, many of the genes in the constructed networks overlap with cancer-related genes, including transcription factors. To confirm this finding, we compiled a list of 496 oncogenes and 874 tumor suppressor genes from the Cancer Genes of Memorial Sloan-Kettering Cancer Center (Higgins et al., 2007) and 1476 human transcription factors (Zhang et al., 2012a). We investigated cancer gene enrichment in the constructed interaction networks by hypergeometric test. As shown in Figure 4.11, most of the significant genes (p-value close to 0) in the constructed networks are overrepresented in the compiled list. This

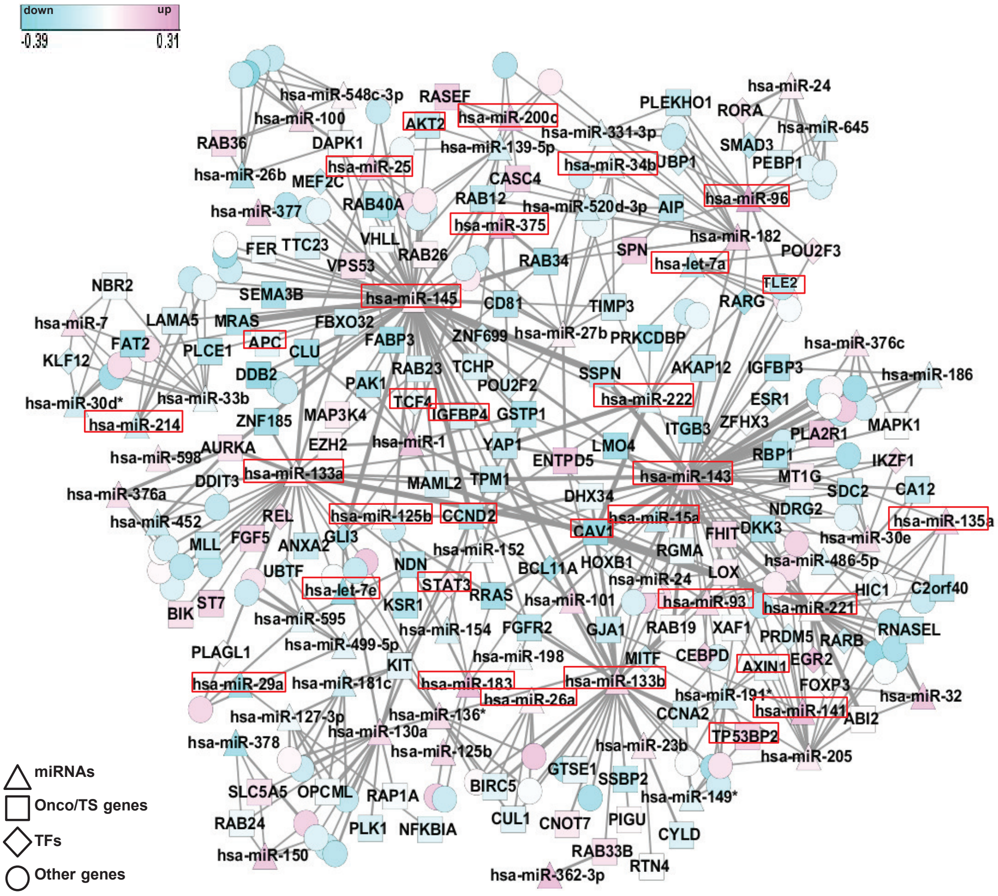


Figure 4.9: Constructed primary prostate cancer-specific miRNA-mRNA interaction networks. The primary-specific network includes 67 miRNAs and 233 mRNAs. The constructed network contains 500 bi-relational edges which are selected based on their summed weight (among all edges converted from 20000 hyperedges of 100 hypergraphs). Up- and down-expressed miRNAs and genes are determined by the mean of each stage class. The red boxed miRNAs and genes have been reported to be associated with the particular stage of prostate cancer. The triangles, rectangles, diamonds and circles denote miRNAs, oncogenes or tumor suppressor genes, transcription factors, and other genes in the network, respectively.

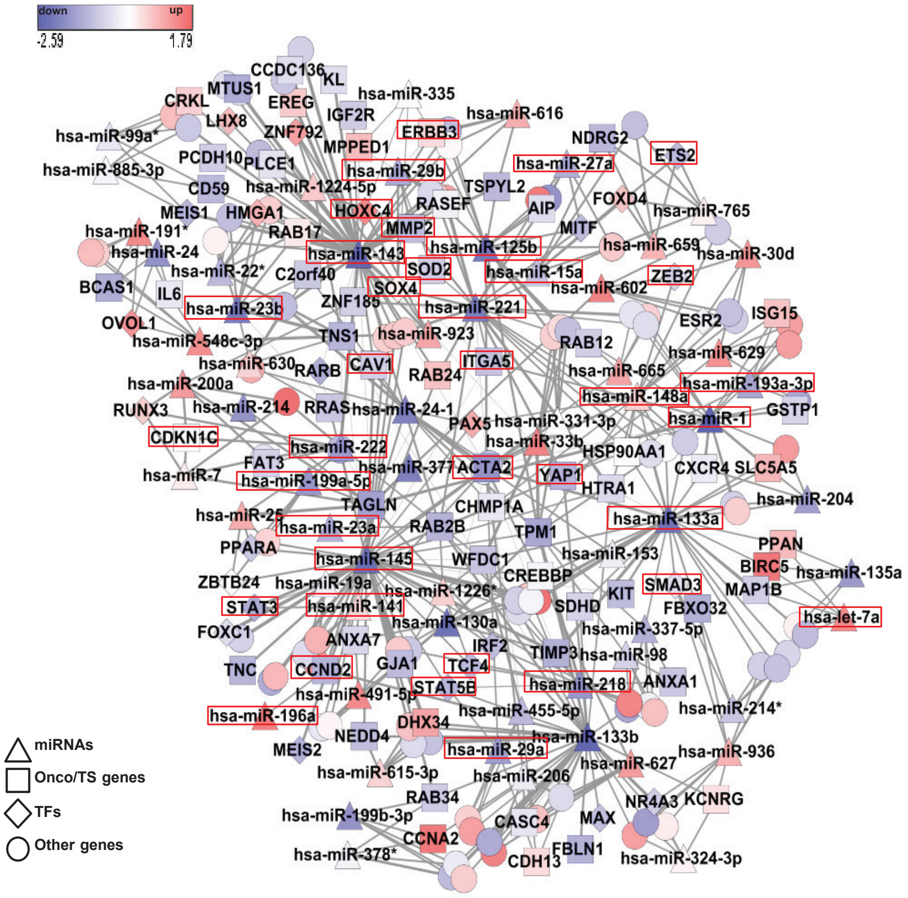


Figure 4.10: Constructed metastatic prostate cancer-specific miRNA-mRNA interaction networks. The metastatic network involves 65 miRNAs and 180 mRNAs. The constructed network includes 500 bi-relational edges which are selected based on their summed weight (among all edges converted from 20000 hyperedges of 100 hypergraphs). Up- and down-expressed miRNAs and genes are determined by the mean of each stage class. The red boxed miRNAs and genes have been reported to be associated with the particular stage of prostate cancer. The triangles, rectangles, diamonds and circles denote miRNAs, oncogenes or tumor suppressor genes, transcription factors, and other genes in the network, respectively.

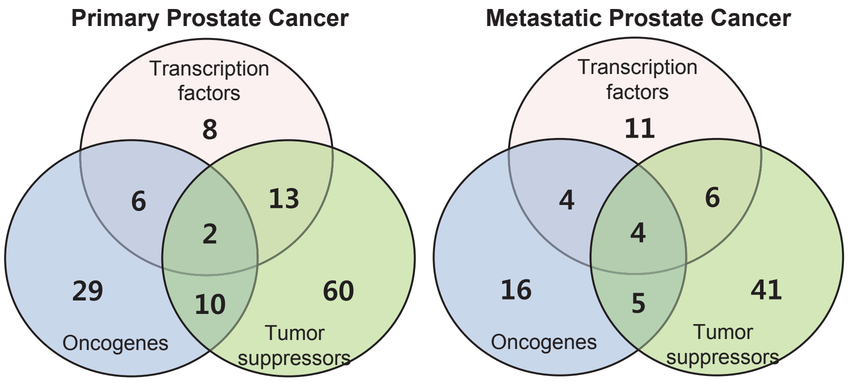
result unambiguously demonstrates that our model can build interaction networks of genetic factors associated with cancer processes.

Interestingly, the enriched hyperedges, and the expression levels of the miRNAs and mRNAs, differ considerably between the primary and metastatic networks. Up- and down-expressed miRNAs and genes are determined by their means at each stage. The red boxed miRNAs and genes are known to be associated with the various stages of prostate cancer (Coppola et al., 2010; Schaefer et al., 2010; Watahiki et al., 2011; Dasgupta et al., 2012; Gordanpour et al., 2012; Triulzi et al., 2013). The triangles rectangles, diamonds and circles denote miRNAs, oncogenes/ tumor suppressor genes, transcription factors, and other genes in the network, respectively.

4.4.5 Functional Analysis of the Constructed Interaction Networks

The constructed miRNA-mRNA interaction networks were validated by functional analyses based on a literature review and gene set analysis. As mentioned above, many of the miRNAs and mRNAs involved in the identified interactions are known indicators of prostate cancer (Coppola et al., 2010; Gordanpour et al., 2012; Watahiki et al., 2011; Schaefer et al., 2010). In addition, the mRNAs comprise a portion of their predicted target genes (Betel et al., 2010), some of which have been experimentally validated. In particular, several miRNAs are known as 'oncomiRs' which function as oncogenes or tumor suppressors, including has-miR-1, -133a, -143, -145, -221, and -222 (Esquela-Kerscher and Slack, 2006; Kojima et al., 2011; Peng et al., 2011; Galardi et al., 2007). Many hyperedges in the constructed networks contain the above miRNAs as their components; these particular miRNAs also act as hubs in the networks.

Especially, hsa-miR-143 and hsa-miR-145 play a crucial role in metastatic prostate



Category (# total genes)	# genes in the network	*p*-value
Primary Prostate cancer		
miRNAs (96)	**28/96**	**4.06e-4**
Transcription factors (1476)	**29/1476**	**2.41e-3**
Oncogenes (495)	**47/495**	**< 0.00e-6**
Tumor suppressor genes (873)	**85/873**	**< 0.00e-6**
Metastatic Prostate cancer		
miRNAs(96)	**23/96**	**1.92e-2**
Transcription factors (1476)	**25/1476**	**8.83e-4**
Oncogenes (495)	**29/495**	**2.22e-16**
Tumor suppressor genes (873)	**56/873**	**< 0.00e-6**

Figure 4.11: The miRNAs and mRNAs in the constructed networks are enriched in cancer-related genes with a significant *p*-value

cancer, and are recognized as a clinicopathological signature of prostate cancer (Peng et al., 2011). Interaction modules involving hsa-miR-143 and -145 occupy a large portion of the networks constructed by our model. In addition, the identified interactions in metastatic prostate cancer contain several experimentally confirmed targets of hsa-miR-143 and -145, including CLINT1, CDKN1A, IRS1, MAPK7, PPM1D and SOD2. Furthermore, hsa-miR-143 and -145 are expressed at low levels in the metastatic network, as has been experimentally validated (Watahiki et al., 2011). Moreover, hsa-miR-200c emerges as a distinct miRNA in the network of primary prostate cancer. According to several studies, hsa-miR-200c overexpression inhibits metastasis prostate cancer, while aberrant regulation triggers the invasion and migration of prostate cancer at the post-transcriptional level (Vrba et al., 2010).

Our model identified several transcription factors associated with prostate cancer metastasis, such as ETS2, HOXC4, STAT3, STAT5B, SOX4 and ZEB2. Among these, SOX4, STAT3 and STAT5B are known regulators of metastatic prostate cancer through the regulation of genes involved in miRNA processing, transcriptional regulation, and developmental pathways (Scharer et al., 2009; Abdulghani et al., 2008; Gu et al., 2010). Indeed, SOX4 is directly regulated by hsa-miR-335 in cancer progression (Scharer et al., 2009), while hsa-miR-125b coordinates STAT3 regulation in the proliferation of tumor cells (Abdulghani et al., 2008).

Interactions involving hsa-miR-29b/MMP2 and hsa-miR-335/SOX4 appear concurrently in the constructed metastatic network (Table 4.3 and 4.4). This finding is consistent with previous studies, in which-miR-29b and -335 were found to suppress tumor metastasis and migration by regulating MMP2 and SOX4, respectively (Triulzi et al., 2013; Steele et al., 2010). Interestingly, both of these interactions involve hsa-miR-143, which is closely linked to prostate cancer progression. Furthermore, the well-known cancer-associated genetic factors MMP2 and SOX4 co-emerged in the identified interactions. Although the interactions identified by our model have not been previously reported, they clearly reflect higher-order relationships between miRNAs and mRNAs. As such, they may signify unknown regulatory circuits in prostate cancer development and progression. This result suggests the utility of the proposed model in identifying undiscovered miRNA-mRNA interactions.

To confirm the biological relevance of the constructed interaction networks, we analyzed the functional correlations among the network genes by canonical path-

Table 4.3: Examples of miRNA-mRNA modules (hyperedges) in primary prostate cancer

♯	miRNA and mRNA modules
1	[miR-330, miR-133$b^{1,2}$, miR-222^{1,3}, MAP1B, WWC3, *CAV1*^{6}, DHX35, TSHZ3]
2	[miR-143^{1,4}, miR-502, miR-548c, ZZEF1, C20orf194, TSPYL2, MBD3, GPR132]
3	[miR-19a^{1}, miR-133$a^{1,2}$, miR-153, BMPR1B, WWC3, PCBP4, TCEAL4, CUL4A]
4	[miR-130a, miR-375, miR-19a^{1}, RAP1A, SNORA71D, CYLD, NDUFA6, RGS9BP]
5	[miR-222^{1,3}, miR-106b, miR-222^{1,3}, ARSJ, SSPN, C3orf58, PTGDS, RARB]
6	[miR-130a, miR-133$a^{1,2}$, miR-19a^{1}, VNN1, FGF5, ELOVL7, PHPT1, RND3]
7	[miR-133$a^{1,2}$, miR-222^{1,3}, miR-130a, SCRIB, FAM108C1, EDRF1, CAR, MOXD1]
8	[miR-130a, miR-149*, miR-26a, RASEF, TPM1, CRB2, GBP, LIX1L]
9	[miR-133$b^{1,2}$, miR-23b, miR-106b, PFAS, UNC5C, HLF, PSEN1, EZH2]
10	[miR-145^{1,4}, miR-200c^{5}, miR-23b, TTC23, PARM1, TOPORS, NEBL, RCAN2]

The underlined genes are the cancer genes archived in the Memorial Sloan-Kettering Cancer Center (MSKCC)^{7}. In addition, genes with a superscript number are confirmed to be related to cancer by the following literature:[1] Esquela-Kerscher and Slack, 2006, [2] Kojima et al., 2012, [3] Galardi et al., 2007, [4] Peng et al., 2011, [5] Vrba et al., 2010, [6] Kypta et al., 2012 and [7] Higgins et al., 2012.

way analysis (Liberzon et al., 2011). The significant (low p-value) results of the analysis for the primary and metastatic prostate cancer networks are summarized in Table 4.5 and 4.6. Many of the enriched pathways are closely associated with prostate tumorigenesis and metastasis. In particular, the β-catenin degradation pathway, the Wnt/β-catenin pathway and the Wnt canonical pathway are associated with Wnt signaling, which regulates many genes implicated in prostate cancer. These pathways were identified as significant in the primary prostate cancer network. Deregulation of the Wnt-related pathway reportedly affects prostate cell

Table 4.4: Examples of modules (hyperedges) in metastatic prostate cancer

♯	miRNA and mRNA modules
1	[miR-221[1,2], miR-29*b*[3], miR-143[1,4,5], *SOX4*[6,8], *MMP2*[3], RASEF, SOD2, SCN9A]
2	[miR-29*b*[3], miR-335[6], miR-143[1,4,5], *SOX4*[6,8], MPPED1, *ERBB3*[9], HOXC4, SMTN]
3	[miR-143[1,4,5], miR-22*, miR-23b, CDKN1A, HMGA1, PELO, RAB17, TMEM150]
4	[miR-125b, miR-616, miR-143[1,4,5], TSPYL2, *ERBB3*[9], ACAD8, PHF15, TMEM16G]
5	[miR-19a, miR-141, miR-145[1,4,5], PCDH20, DNAJC3, *STAT3*[10,11], ZNF385, ACTA2]
6	[miR-133*b*[1,7], miR-145[1,4,5], miR-218, IRF2, *TCF4*[12], *STAT5B*[13], RAB2B, WFDC1]
7	[miR-143[1,4,5], miR-145[1,4,5], miR-222[1,2], ITGA5, MAPK7, MAP3K2, RAB34, S100A1]
8	[miR-214, miR-143[1,4,5], miR-145[1,4,5], FEM1A, ITGA5, NAGPA, C1orf142, ERAS]
9	[miR-193b, miR-143[1,4,5], miR-145[1,4,5], CLINT1, GJA1, MAPK7, RARRES2, IL28A]
10	[miR-221[1,2], miR-1[1,7], miR-133*b*[1,7], *TPM1*[12], NDFIP2, RAD17, VPS28, INPPd5E]

The underlined genes are the cancer genes archived in the Memorial Sloan-Kettering Cancer Center (MSKCC)[14]. In addition, genes with a superscript number are confirmed to be related to cancer by the following literature:[1] Esquela-Kerscher and Slack, 2006, [2] Galardi et al., 2007, [3] Steele et al., 2010, [4] Watahiki et al., 2011, [5] Peng et al., 2011, [6] Triulzi et al., 2013, [7] Kojima et al., 2012, [8] Scharer et al., 2009, [9] Schwaetz et al., 1999, [10] Abdulghani et al., 2008, [11] Haghikia et al., 2012, [12] Kypta et al., 2012, [13] Gu et al., 2010, and [14] Higgins et al., 2012.

proliferation and differentiation (Kypta and Waxman, 2012). Moreover, the annotated genes in the constructed network, such as APC, AXIN1, AKT2, CCND2, CAV1, TLE2 and TCF4, are essential regulatory components of these pathways in prostate cancer. ErbB-related pathways were identified in the metastatic network, including the ErbB network pathway, ErbB4 pathway, Her2 pathway, ErbB2/ErbB3 signaling pathway and the EGFR pathway, which are implicated in prostate cancer progression and metastasis (Dasgupta et al., 2012; Schwartz et al., 1999). The

Table 4.5: Canonical pathway analysis of the constructed interaction networks in primary prostate cancer

Canonical Pathway Analysis	p-value (< 0.05)
Pathways in cancer	1.70e-03
Rb1 pathway	5.95e-03
Retinoic acid pathway	6.61e-03
Aurora A pathway	7.44e-03
Beta-catenin degradation pathway	9.95e-03
Wnt/beta-catenin pathway	1.03e-02
Wnt canonical signaling pathway	1.34e-02
Met pathway (signaling of HGF receptor)	1.39e-02
P38-alpha/beta downstream pathway	1.52e-02
Beta-catenin nuclear pathway	1.58e-02
Aurora B pathway	1.66e-02
EPHB forward pathway	1.81e-02
IFN-gamma pathway	1.81e-02
P53 hypoxia pathway	1.97e-02
MYC repress pathway	2.15e-02
Progesterone mediated oocyte maturation	2.19e-02
Rac CycD pathway (Ras and Rho protein on G1/S transition)	2.73e-02
PLK1 pathway	2.88e-02
IL-6 (interleukin-6) pathway	3.08e-02
FGFR2C ligand binding and activation	3.58e-02
Cell cycle	4.43e-02
PDGFR-beta signaling pathway	4.59e-02

Table 4.6: Canonical pathway analysis of the constructed interaction networks in metastatic prostate cancer

Canonical Pathway Analysis	*p*-value (< 0.05)
MYC activate pathway	1.41e-04
ErbB network pathway	2.78e-03
KIT receptor signaling pathway	3.28e-03
IL-10 pathway	4.40e-03
Pathways in cancer	4.76e-03
ErbB4 pathway	6.12e-03
Her2 pathway (ErbB2 in signal transduction and oncology)	8.51e-03
Yap1 and Wwtr1/Taz stimulated gene expression	1.09e-02
Smooth Muscle Contraction	1.22e-02
Barrestin pathway	1.53e-02
IL-6 signaling pathway	1.85e-02
STAT3 pathway	1.85e-02
IL-2/STAT5 pathway	2.00e-02
RAS pathway	2.00e-02
ErbB2/ErbB3 signaling pathway	2.19e-02
Syndecan4 pathway	2.38e-02
PPAR-alpha pathway	2.61e-02
Integrin signaling pathway	3.72e-02
Rela pathway	3.78e-02
HDAC class I pathway	3.94e-02
FOXM1 pathway	4.24e-02
IL-7 pathway	4.23e-02
EGFR pathway	4.70e-02

FOXM1 pathway also regulates tumor metastasis (including that of prostate cancer) by stimulating the expression of several genes involved in the proliferation of tumor cells and cell cycle progression (Raychaudhuri and Park, 2011). The top-ranked pathway in the metastatic network is the MYC activation pathway. MYC reportedly promotes the metastatic phenotype by altering the epigenetic landscape of cancer cells, and is overexpressed in ~75% of advanced prostate cancer patients (Dasgupta et al., 2012). Thus, the MYC pathway is a putative key feature of metastatic progression (Wolfer and Ramaswamy, 2011).

4.5 Summary

The proposed hypergraph-based model characterizes higher-order interactions among heterogeneous genetic factors from archived data. Human cancers are typically caused by the modular control of multiple genetic factors. By analyzing gene relationships at higher-order levels, thus, we can better understand the behavior of complex cancer mechanisms. Moreover, the cooperative activities and the combinatorial regulations governed by miRNAs and mRNAs are largely unknown. We have demonstrated that higher-order relationships discriminate between specific cancer stages more precisely than pair-wise analyzes of single miRNA and mRNA interactions. From this viewpoint, we can construct a more complete interaction network consisting of putative biologically significant miRNA-mRNA modules.

In addition, our method focuses on discovering potential interactions in unknown miRNA-mRNA regulatory circuits related to specific cancer stages without the known biological information (Friedman, 2004; Ivan et al., 2008). The proposed model finds statistically significant gene modules from given expression profiles using a data-driven approach with co-regulatory measure (mutual information). However, a similar hypergraph structure could be readily constructed from other types of quantitative biological information, such as miRNA-target information and gene sequence similarity values. Furthermore, the hypergraph-based model more flexibly represents miRNA-RNA interactions than other methods (which assume that the expression states of miRNAs and mRNAs are linearly proportional to each other), because it isolates significant modules from the statistical co-expressed pattern among genes at a higher-order level.

The proposed hypergraph model is similar to Bonnet's et al. (Bonnet et al., 2010a,b) and Li et al. (Liu et al., 2010), where higher-order relationships governed by miRNA-mRNA interactions are inferred solely from expression profiles. Bonnet's method is based on a clustering approach, it cannot readily infer gene regulatory modules at a specific cancer stage. In contrast to Bonnet's method, our method explicitly considers the sample status, (the primary or metastatic state of prostate cancer), from which it constructs cancer stage-specific networks. Liu's approach is based on Corr-LDA, which requires that data are discretized. By contrast, our method uses intact real-valued data, thus preventing the information loss caused by the discretization.

In brief, we have proposed a hypergraph-based model consisting of higher-order miRNA-mRNA modules, which allows the construction of biologically meaningful interaction networks associated with specific cancer stages. For identifying potential significant interactions and refining model performance, we introduced a two-phase learning approach comprising structure and parameter learning. Finally, we constructed cancer stage-specific interaction networks reflecting higher-order miRNA and mRNA relationships by converting the hypergraph structure into an ordinary graph.

We constructed higher-order miRNA-mRNA interaction networks associated with the specific stage of prostate cancer from a matched dataset using the proposed model. The performance of the proposed model is similar to that of SVMs and superior to other classification models (outperforming them by approximately 6-10 %). More importantly, our model can construct carcinogenic miRNA-hubbed networks that characterize primary and metastatic prostate cancer. Furthermore, we demonstrated that a large proportion of the miRNAs and mRNAs identified in the constructed interaction networks are indeed involved in prostate cancer progression and development. The proposed hypergraph-based model therefore presents as an alternative method for discovering potential gene regulatory circuits. Such discoveries will greatly assist our understanding of cancer pathogenesis.

Chapter 5

Hierarchical Hypergraphs for Identifying Higher-Order Genomic Interactions in Multilevel Regulation

5.1 Overview

The importance of epigenetics has been increasingly recognized in various biological processes. Epigenetic mechanisms play important roles in controlling and maintaining normal gene expression pattern via modification or rearrangement of nucleosomes by changing the accessibility of chromatin to transcriptional regulation (Bonetta, 2008). Especially, DNA methylation is a crucial epigenetic regulation in various diseases pathogenesis including carcinogenesis (Esteller, 2007; Jones, 2012). DNA methylation typically occurs at CpG islands of promoters by DNA methyltransferase (DNMT) enzyme without DNA sequence alterations, and affects transcriptional behavior in cells such as gene silencing and activating (Laird, 2010). Aberrant DNA methylation contributes to the malignant phenotype of human cancer cells as a hallmark of tumorigenesis. While cooperating with genetic alterations, also, epigenetic regulation including DNA methylation is strongly implicated in tumor initiation, development, proliferation, and suppression (Egger et al., 2004; Jones and Baylin, 2007; Handel et al., 2010). Thus, the combinatorial analysis between epigenetic and genetic factors is necessary to understand complex cancer mechanisms at the molecular level.

Ovarian cancer is one of the most deadly gynecological malignancy in the world, caused by combinatorial effects of multiple factors (Jemal et al., 2010). Abnormal DNA methylation is a common phenomenon in ovarian cancer, and closely associated with the initiation and progression of ovarian cancer by regulating multiple genetic factors such as microRNAs (miRNAs) and mRNAs (Holschneider and Berek, 2000). Herein, the coordinated regulation of miRNAs and mRNAs involved in DNA methylation should be elucidated to systemically explore the mechanism of ovarian cancer.

Here, we propose a hierarchical hypergraph model to identify higher-order miRNA-mRNA interactions associated with the regulation of DNA methylation from TCGA data (Figure 5.1). The proposed model explicitly characterizes complex relationships among multiple genomic factors involved in the specific epigenetic regulation, from which correlated gene interactions between methylome and transcriptome in biological processes including cancer pathogenesis may be identified. A hierarchy is introduced into the hypergraph model by defining two layers representing each epigenetic and genetic regulation level. The first layer consists of hyperedges that encode higher-order relationships among many genomic factors same as the traditional hypergraphs. And the second layer is composed of vari-

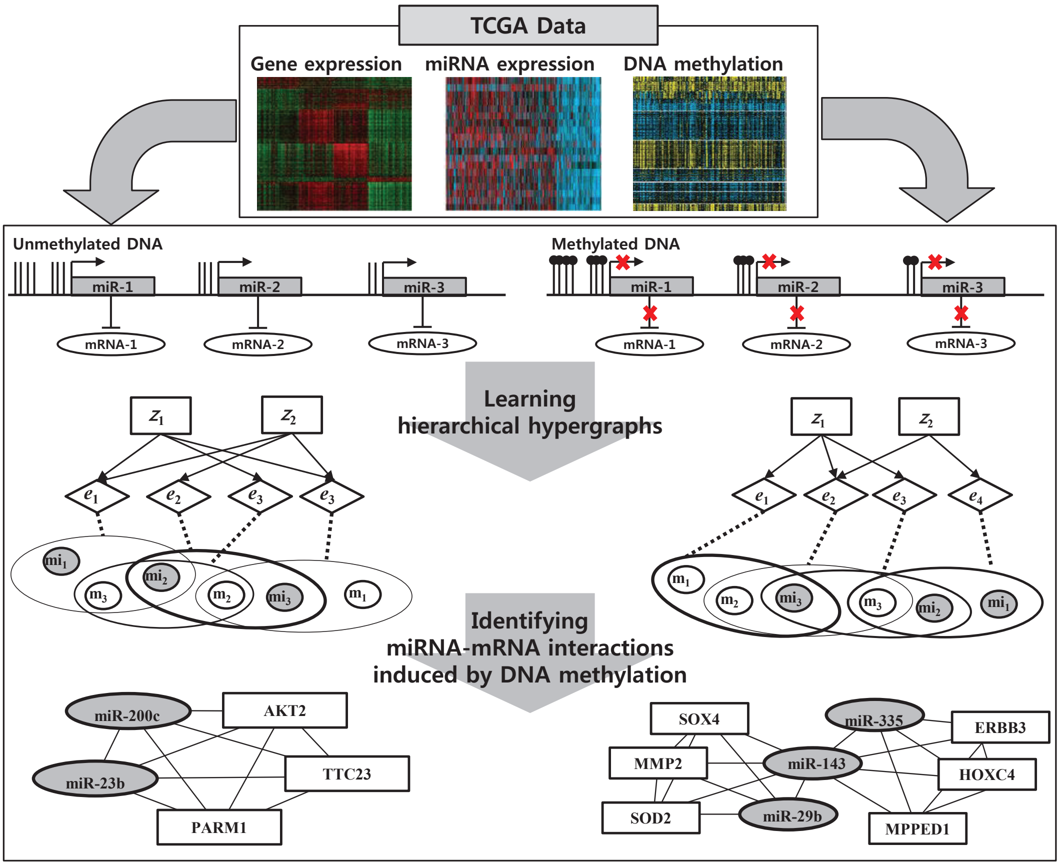


Figure 5.1: Overview of the hierarchical hypergraph for identifying higher-order genomic interactions induced by the specific DNA methylation regulation.

ables characterizing biological function and regulation. The learning of hierarchical hypergraphs proceeds by repeating three steps: generating hyperedges, calculating the objective function, and removing hyperedges with low weight. This learning method is designed upon a standard evolutionary computation framework.

The goal of the learning is to identify significant DNA methylation changes underlying cancer, and miRNA-mRNA regulatory interactions induced by the methylation change. For achieving this goal, we define an objective function which reflects the strength of interactions between miRNAs and mRNAs associated with the specific DNA methylation events from multisource genomic data using information theoretic co-regulatory measure, called mutual information. Moreover, the higher-order relationships among genomic variables are intractably complex, we adopt an evolutionary strategy for efficient searching. This hierarchical structure learning allows the model to detect potential gene regulatory circuits across the level of epigenetic, transcriptional and post-transcriptional regulation.

We identify higher-order miRNA-mRNA interactions involved in specific DNA methylation changes in ovarian cancer using TCGA data (Bell et al., 2011) from the model. We demonstrate that the proposed model can find several biologically significant miRNA-mRNA interactions implicated in DNA methylation regulation, including potential modules associated with ovarian cancer. Moreover, cancer-related miRNAs and genes dominate the identified interactions. We also confirm the biological significance of the identified interactions through literature review and functional analysis.

5.2 Analyzing Epigenetic and Genetic Interactions from Multiple Genomic Data

Recent epigenetic research has progressed to obtain a global view of gene regulation at multi-cellular level. Many studies have focused on analyzing the relationships between only two data sources, such as DNA methylation-genes (Siegfried and Simon, 2010; Spisák et al., 2012; van Eijk et al., 2012; Busche et al., 2013; Marx et al., 2013) and DNA methylation-miRNAs (Han et al., 2007; Lujambio et al., 2008;

Yan et al., 2011; Baer et al., 2012; Wong et al., 2012), on genome-wide scale from high-throughput data. These approaches have systemically investigated the complex mechanism of various cancers at the multi-level regulation. However, it is difficult to directly extract the regulatory modules between epigenetic and genetic components underlying specific cancer types. To overcome this issue, Joung et al. (Joung et al., 2013) proposed a method to extract the correlated gene pairs to DNA methylation from both expression profiles. This method calculates the unified score, consisting of the differential score and correlated score, which measures the strength of the regulatory relationships between two genes. However, the score reflects the pairwise relations of only two genes, and thus this method is difficult to precisely address complex genetic interactions associated with the epigenetic events. In recent, moreover, several studies have attempted to simultaneously explore the coordinated relationships from heterogeneous data, such as DNA methylation, gene and miRNA expression profiles, via computational approaches including statistical method (Zhu et al., 2011), matrix factorization (Zhang et al., 2012b) and regression model (Li et al., 2012). However, analyzing higher-order relationships across the levels of epigenetic, transcriptional, and post-transcriptional regulations is still rendered as a challenging issue due to the complexity of their interactions.

5.3 Hierarchical Hypergraphs for Identifying Epigenetic and Genetic Interactions

5.3.1 Hierarchical Hypergraphs

Hierarchical hypergraphs is a hypergraph-based model consisting of two distinct layers. The first layer includes hyperedges whose nodes are observable target variables while latent or observable causal variables exist in the second layer. The

理學博士學位論文

Hypergraph Models for Identifying Co-Regulatory Genomic Interactions

동시조절 유전적 상호작용 발굴을 위한 하이퍼그래프 모델

2014年 2月

서울大學校 大學院

협동과정 생물정보학

金 秀 珍

Hypergraph Models for Identifying Co-Regulatory Genomic Interactions

동시조절 유전적 상호작용 발굴을 위한 하이퍼그래프 모델

指導敎授　張 炳 卓

이 論文을 理學博士 學位論文으로 提出함

2013年 10月

서울大學校 大學院

협동과정 생물정보학

金 秀 珍

金秀珍의 理學博士 學位論文을 認准함

2013年 11月

委 員 長 李 柄 宰

副委員長 張 炳 卓

委　　員 尹 晟 老

委　　員 鄭 東 均

委　　員 申 守 容

Hypergraph Models for Identifying Co-Regulatory Genomic Interactions

동시조절 유전적 상호작용 발굴을 위한 하이퍼그래프 모델

Soo-Jin Kim

Ph.D. Thesis

Interdisciplinary Program in Bioinformatics
Seoul National University

February 2014

Supervisor: Byoung-Tak Zhang

Abstract

A comprehensive understanding of biological systems requires the analysis of higher-order interactions among many genomic factors. Various genomic factors cooperate to affect biological processes including cancer occurrence, progression and metastasis. However, the complexity of genomic interactions presents a major barrier to identifying their co-regulatory roles and functional effects. Thus, this dissertation addresses the problem of analyzing complex relationships among many genomic factors in biological processes including cancers. We propose a hypergraph approach for modeling, learning and extracting: explicitly modeling higher-order genomic interactions, efficiently learning based on evolutionary methods, and effectively extracting biological knowledge from the model.

A hypergraph model is a higher-order graphical model explicitly representing complex relationships among many variables from high-dimensional data. This property allows the proposed model to be suitable for the analysis of biological and medical phenomena characterizing higher-order interactions between various genomic factors. This dissertation proposes the advanced hypergraph-based models in terms of the learning methods and the model structures to analyze large-scale biological data focusing on identifying co-regulatory genomic interactions on a genome-wide level. We introduce an evolutionary approach based on information-theoretic criteria into the learning mechanisms for efficiently searching a huge problem space reflecting higher-order interactions between factors. This evolutionary learning is explained from the perspective of a sequential Bayesian sampling framework. Also, a hierarchy is introduced into the hypergraph model for modeling hierarchical genomic relationships. This hierarchical structure allows the hypergraph model to explicitly represent gene regulatory circuits as functional blocks or groups across the level of epigenetic, transcriptional, and post-transcriptional regulation. Moreover, the proposed graph-analyzing method is able to grasp the global structures of biological systems such as genomic modules and regulatory networks by analyzing the learned model structures.

The proposed model is applied to analyzing cancer genomics considered as a major topic in current biology and medicine. We show that the performance of our model competes with or outperforms state-of-the-art models on multiple cancer genomic data. Furthermore, the propose model is capable of discovering new or hidden patterns as candidates of potential gene regulatory circuits such as gene modules, miRNA-mRNA networks, and multiple genomic interactions, associated with the specific cancer. The results of these analysis can provide several crucial evidences that can pave the way for identifying unknown functions in the cancer system. The proposed hypergraph model will contribute to elucidating core regulatory mechanisms and to comprehensive understanding of biological processes including cancers.

**Keywords: Hypergraph, Higher-order graphical model,
Evolutionary learning, Genomic interaction,
Gene module, miRNA-mRNA network, Cancer**

Student Number: 2005-20561

Contents

List of Tables

List of Figures

Chapter 1

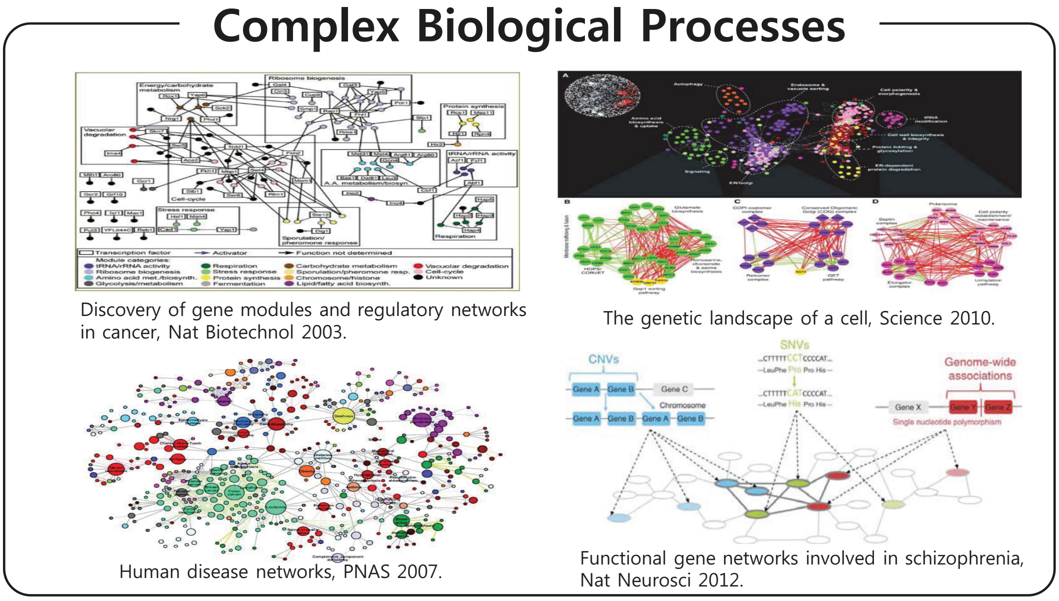
Introduction

1.1 Background and Motivation

Recent biological and medical research advances from studying simple to complex traits, from carrying out separated to integrated analyzes of different genomic data sources, from analyzing a single gene to multiple gene interactions at the systems level. Computational approaches, which analyze gene regulatory relationships on a genome-wide scale from high-throughput data, have led to a deluge of systemic insights into a variety of biological and medical areas.

Cancer is one of the important challenges in biology and medicine because it is still the lethal disease of the leading cause of death worldwide. High-throughput data have been massively produced to understand cancer mechanisms for more several years. Despite such efforts, the mechanism of cancer is not clearly deciphered yet.

The regulation of cancer is a complicated phenomenon, induced by complex interactions among various genetic factors. It is mostly related to modular construction and combinatorial control by multiple genetic factors such as miRNAs



Complex Biological Processes

Discovery of gene modules and regulatory networks in cancer, Nat Biotechnol 2003.

The genetic landscape of a cell, Science 2010.

Human disease networks, PNAS 2007.

Functional gene networks involved in schizophrenia, Nat Neurosci 2012.

Figure 1.1: Analyzing higher-order genomic interactions is necessary to understand complex and various biological processes including cancer.

and mRNAs across the transcriptional, post-transcriptional and epigenetic levels. Thus, elucidating multiple genomic interactions at multicellular level is essential to understand complex biological processes including cancer development and progression more precisely (Figure 1.1). Furthermore, it can provide new insights into the behavior of complex biological systems. However, the analysis of higher-order relationships between many genetic factors is rendered as a challenging problem due to the complexity of their interactions.

Herein, one of the major issues associated with investigating complex genomic interactions is the volume of data to be analyzed; as the number of genes increases the number of potential interactions increases exponentially, known as 'curse of dimensionality' (Moore and Ritchie, 2004). The potential complexity of such interactions supports the use of various machine learning techniques for analyzing co-regulatory genomic relationships implicated in complex diseases including cancer.

Over the past ten or more years, many models and algorithms based on machine learning have been developed for analyzing complex biological systems, and contributed to rapid advances in biology and medicine by providing new solutions (Larranaga et al., 2006; McKinney et al., 2006; Fogel, 2008; Upstill-Goddard et al., 2013). Because biological systems are inherently non-linear and dynamic, the proper comprehension of such systems requires interpretive methods that do not rely strictly on linearity and can deal with complex relationships.

Higher-order models represent complex interactions among many factors with higher-order units as their features instead of data variables or linear summations of the variables (Roddick et al., 2008). Higher-order models can more precisely characterize the complicated dependencies embodied in biological phenomena, thus providing better modeling performance than simple linear models (Lehar et al., 2008). Such models based on higher-order representations can be complementary to existing approaches, and can be used to search very large solution spaces efficiently for analyzing complex biological processes including cancer. The range of recent successful applications makes it all the more evident that the need for these models will continue to increase in the near future.

1.2 Problems to be Addressed

Many real-world problems in biological and medical fields require higher-order representation of complex dependency among various factors. Moreover, recent advances in high-throughput molecular techniques have resulted in the exponential growth of the amount of biological data that reflect the interplay between biomolecules on a genome-wide scale. Due to the complexity of the regulatory mechanisms involved and the large number of possible interactions, it is a great need for computational approaches which enable to systematically and efficiently analyze complex biological processes.

This dissertation proposes a higher-order graphical model for dealing with complex relations between many factors and focuses on analyzing co-regulatory genomic interactions on a genome-wide scale for understanding various biological processes including cancers with the new proposed higher-order model. The proposed model can naturally learn the higher-order patterns from high-dimensional data by the process of selecting hyperedges and adjusting their weights. However, since the number of possible hyperedges grows exponentially with the number of features and their combinations, it results in a need for effective learning strategies and suitable model structures to solve complex biological problems. Moreover, it requires the method for extracting meaningful biological knowledge from the learned model.

In this dissertation, we mainly addressed three issues: 1) the advanced model structure for representing higher-order interactions between numerous genomic factors, 2) the improved learning method to efficiently search huge combinatorial feature spaces from very high-dimensional biological data, and 3) the novel method for extracting meaningful biological knowledge from the learned models.

1.3 The Proposed Approach and its Contribution

We propose the advanced class of higher-order graphical models for analyzing complex biological problems incurred by the large number of higher-order interactions, and the improved learning method for efficiently searching huge problem spaces from high-dimensional data. In addition, a novel graph-analyzing method is proposed to extract meaningful biological information and knowledge from the learned models.

The proposed model structure explicitly characterizes higher-order interactions among numerous genomic factors, from which cooperative gene activities in biological processes may be identified. It adopts a flexible hypergraph structure composed of a large population of hyperedges, representing the multi-variable combinations consisting of a variety of genomic factors. Thus, the structure of the proposed model is effective to represent higher-order genomic interactions or complex gene modules for analyzing co-regulatory gene mechanisms in various biological processes including cancer.

The learning of hypergraph models involves searching a huge combinatorial feature space due to its definition and the problem space exponentially enlarges as the number of features increase. This issue becomes more severe when applied to large-scale biological data which consists of several tens of thousands variables. The proposed learning method is able to efficiently search a huge problem space reflecting higher-order relationships between factors by introducing information-theoretical criteria for a guided search into the conventional evolutionary learning approach. The proposed learning method is explained with a sequential Bayesian sampling framework.

Finally, this dissertation proposes a method to enable identify co-regulatory gene modules or to construct gene regulatory networks from the learned higher-order model. Although it is important to extract meaningful information and knowledge from the models in biological and medical fields, the previous studies on hypergraph models focused on the learning efficiency and the model performance rather than knowledge extraction by analyzing the learned model. A network characteriz-

The advanced hypergraph-based models

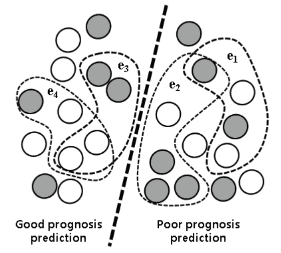
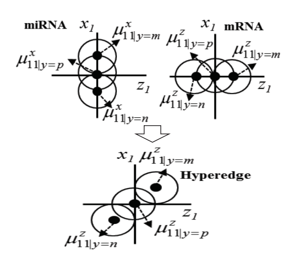
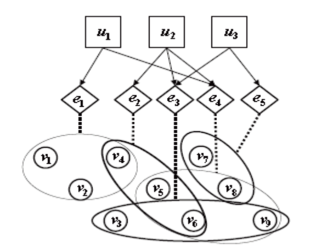
Model	Hypergraph classifiers for discrete variable data	Hypergraph models for continuous variable data	Hierarchical hypergraphs for multiple interactions
			
Property	• **Standard model** • Single source analysis • Discrete-valued data process	• **Generalized representation** • Multi-source analysis • **Real-valued data process**	• **Advanced structure** • **Hierarchical relations analysis** • Real-valued data process

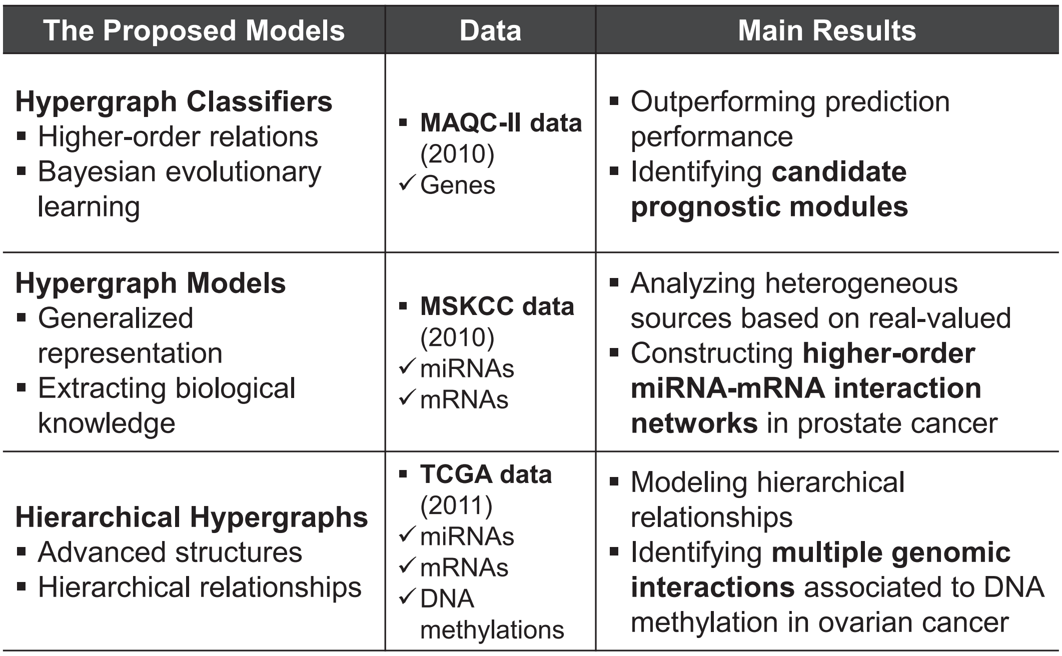
Figure 1.2: The improvement of the proposed models in this dissertation.

ing higher-order genomic interactions is constructed from the leaned hypergraphs based on a minimum-cut approach in this dissertation. Thus, the proposed model can directly extract meaningful knowledge such as co-regulatory gene modules, pathways or networks from various genomic data. Furthermore, it can discover new or potential genomic regulatory circuits which assist our understanding of biological systems including cancer pathogenesis.

Figure 1.2 illustrates the improvement of the proposed models in this dissertation and Figure 1.3 summarized the main results by the proposed model.

1.4 Organization of the Dissertation

This dissertation is organized as follows:



The Proposed Models	Data	Main Results
Hypergraph Classifiers ▪ Higher-order relations ▪ Bayesian evolutionary learning	▪ **MAQC-II data** (2010) ✓Genes	▪ Outperforming prediction performance ▪ Identifying **candidate prognostic modules**
Hypergraph Models ▪ Generalized representation ▪ Extracting biological knowledge	▪ **MSKCC data** (2010) ✓miRNAs ✓mRNAs	▪ Analyzing heterogeneous sources based on real-valued ▪ Constructing **higher-order miRNA-mRNA interaction networks** in prostate cancer
Hierarchical Hypergraphs ▪ Advanced structures ▪ Hierarchical relationships	▪ **TCGA data** (2011) ✓miRNAs ✓mRNAs ✓DNA methylations	▪ Modeling hierarchical relationships ▪ Identifying **multiple genomic interactions** associated to DNA methylation in ovarian cancer

Figure 1.3: Main results by the proposed models

- Chapter 2 presents a survey of the related work. Firstly, we discuss the previous research on the analysis of co-regulatory gene interactions in genomes. Also, we summarize probabilistic graphical models including Bayesian networks, Markov random fields, and hidden Markov models for biological problems. Next, we explain the concept of higher-order model, and summarize previous studies on higher-order graphical models including hypergraphs. In addition, we introduce the applications of hypergraphs and hypergraph-based models in biological problems.

- In Chapter 3, we propose hypergraph classifiers to identify prognostic gene modules for predicting cancer clinical outcomes. The proposed hypergraph classifier is based on evolutionary learning that identifies higher-order gene modules of cancer clinical outcomes. We demonstrate our model can deal with high dimensional data more effectively than state-of-the-art classification models, and identify potential gene modules characterizing prognosis and recurrence risk in cancer.

- Chapter 4 describes the advanced hypergraph model for identifying higher-order genomic interactions from heterogeneous. And we suggest a method for constructing interpretable networks reflecting such higher-order interactions from the learned hypergraph model. We show that the proposed model can build higher-order miRNA-mRNA interaction networks using MSKCC prostate oncogenome data. Also we confirm the biological relevance of the constructed networks through literature review and functional analysis.

- In Chapter 5, we introduce a hierarchical hypergraph model to identify multiple genomic interactions involved in the specific epigenetic mechanisms. A hierarchy are introduced into the hypergraph model by defining two layers. This hierarchical structure allows the proposed model to analyze higher-order genomic relationships at the multi-level regulation. We demonstrate our model can identify higher-order miRNA-mRNA interactions involved in the specific DNA methylation regulation on a genome-wide scale from TCGA data.

- This dissertation is summarized and directions for further research are discussed in Chapter 6.

Chapter 2

Related Work

2.1 Analysis of Co-Regulatory Genomic Interactions from Omics Data

The availability of high-throughput omics data have opened up a new possibility to study the interaction of genetic components underlying the specific biological process such as tumorigenesis at the systems level. Rapid advances in computational approaches which analyze such large-scale data offer a new conceptual framework that can potential revolutionize our view of biology and disease pathologies.

Several years ago, B. Alberts and L. Hartwell noted that biological processes are organized into co-regulatory groups or modules, and that the reductionist approach for studying each process in isolation is limiting (Alberts, 1998; Hartwell et al., 1999). For this reason, one of key issues in current computational systems biology is to systematically analyze gene regulatory mechanisms by using module-based approach from various omics data. Many efforts have taken advantage of this view to investigate a variety of biological processes such as cancer onset, progression and metastasis, which consist of complex interactions among many genetic components on a genome-wide scale (Segal et al., 2001, 2003, 2004, 2005; Bonneau, 2008; Barabási et al., 2011).

Modern cancer research has progressed from identifying biomarkers to systemically exploring gene interactions (Hornberg et al., 2006; Wang et al., 2007; Liu et al., 2012). Many studies have attempted to describe how genetic components interact on the system level. Computational methods, which analyze gene regulatory networks and interactions on a genome-wide scale from high-throughput biological data, have flourished in recent decades (Bar-Joseph et al., 2003; Schlitt and Brazma, 2007; Yan et al., 2007; Lee and Tzou, 2009; Joung et al., 2012; Mitra et al., 2013). In addition, systems biology approaches to study miRNA regulation were designed to understand the development of multiple human malignancies (Kim et al., 2006; Bandyopadhyay et al., 2010; Volinia et al., 2010). Moreover, recent studies have focused on reconstructing regulatory networks by integrating miRNAs and other molecules such as mRNAs, transcriptional factors, and proteins for different physiological and pathological conditions (Shalgi et al., 2007; Bonnet et al., 2010a; Nasser et al., 2010; Li et al., 2011; Lu et al., 2010; Zhang et al., 2011).

Those approaches have helped to simplify complex biological mechanisms by systemically analyzing the relationships between genetic elements at the genome level. However, many studies on this issue use an approach considering relationships between only two factors for analyzing the interactions among genes. In addition, we are still far from understanding the mechanisms of cooperative regulations among various components in a specific biological process. Therefore, inferring regulatory networks by taking into consideration the complex dependencies among genetic factors remains a formidable challenge.

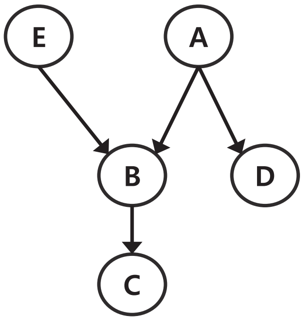
2.2 Probabilistic Graphical Models for Biological Problems

Probabilistic graphical models (PGMs) have been applied to many real-world problems. The general framework of PGMs uses ideas from discrete data structures in computer science to efficiently encode and manipulate probability distributions over high-dimensional spaces, often involving hundreds or even many thousands of variables (Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012). These models have been used in an enormous range of application domains, which include: medical diagnosis, biological network reconstruction, speech recognition, natural language processing, intelligent control, and many more.

2.2.1 Bayesian Networks

A Bayesian network (Heckerman et al., 1995; Jensen, 1996; Friedman et al., 1997; Neapolitan, 2004; Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012) is a graphical model that encodes probabilistic relationships among variables of interest. This model is more suitable for analyzing biological data because it can represent cause and effect relationships. Graphical model has several advantages for data analysis when used in conjunction with Bayesian statistical techniques. Firstly, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Second, Bayesian networks can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Third, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data.

Bayesian networks compactly represent the joint probability distribution over a set of random variables via a directed acyclic graph (DAG). In the framework of probabilistic graphical model, the concept of conditional independence is exploited



Conditional independence statements:
I(A; E), I(B; D | A, E), I(C; A, D, E | B),
I(D; B, C, E | A), and I (E; A, D)

Product form of joint distribution:
P(A, B, C, D, E) = P(A)P(B|A, E)P(C|B)P(D|A)P(E)

Figure 2.1: An example of a simple Bayesian network structure

for efficient representation of joint probability distribution. For three variable sets $\{X, Y, Z\}$, X is conditionally independent from Y given the value of Z. The Bayesian network structure encodes various conditional independencies among the variables as follows. A Bayesian network assumes a directed acyclic graph structure where each node corresponds to a variable and an edge is a direct probabilistic dependency between the two connected nodes. Formally, the DAG structure asserts that each node is independent of all its non-descendants conditioned on its parent nodes. A Bayesian network consisting of n variables, $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, represents a joint probability distribution $\mathbf{P}(\mathbf{X})$ given by the product of all conditional probability:

$$\mathbf{P}(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | pa^G(X_i)) \tag{2.1}$$

where $pa^G(X_i)$ is the set of parents of X_i in the DAG structure G, and $\mathbf{P}(\mathbf{X})$ reflects the properties of Bayesian network. A graph G specifies a product form as in Figure 2.1. To fully specify a joint distribution, we also need to specify each of the conditional probabilities in the product form. The second part of the Bayesian network describes these conditional distributions, $P(X_i | pa^G(X_i))$ for each variable

X_i.

Bayesian network structure learned from data can provide us insight into the complicated cause and effect relationship among a set of variables. Thus, it is applicable for extracting knowledge from data. As such, Bayesian network has been widely applied in various areas including cancer diagnosis (Nikovski, 2000; Gevaert et al., 2006; Cruz-Ramírez et al., 2007) and gene expression analysis (Friedman et al., 2000; Segal et al., 2003; Imoto et al., 2004; Liu et al., 2013).

2.2.2 Markov Random Fields

A Markov random field (MRF) (Kindermann and Snell, 1980; Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012) also known as a Markov network, or a probabilistic independence network, is a set of random variables having a Markov property described by an undirected graph (Figure 2.2 (a)). Every variable X_i is represented by a node in the graph and the nodes are connected by undirected edges. Let $adj(X_i)$ be all the nodes that are adjacent (i.e., directly connected) to X_i, then the edges in a Markov field are places in such way that:

$$\forall X_j \in \chi \backslash X_i \cup adj(X_i); X_i \perp\!\!\!\perp X_j | adj(X_i) \tag{2.2}$$

where χ is the value set of X_i and $adj(X_i)$ acts as the Markov blanket of X_i. In contrast to belief networks, there are no conditional probability functions connected to nodes. Instead, each *clique* in the graph is provided with a *potential* $\psi_c(\cdot)$ which assigns a non-negative real value to all combinations of values of nodes. A clique is defined as a subset of the nodes in a graph such that there exists a link between all pairs of nodes in the subset. In other words, the set of nodes in a clique is fully connected. Furthermore, a *maximal clique* is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.

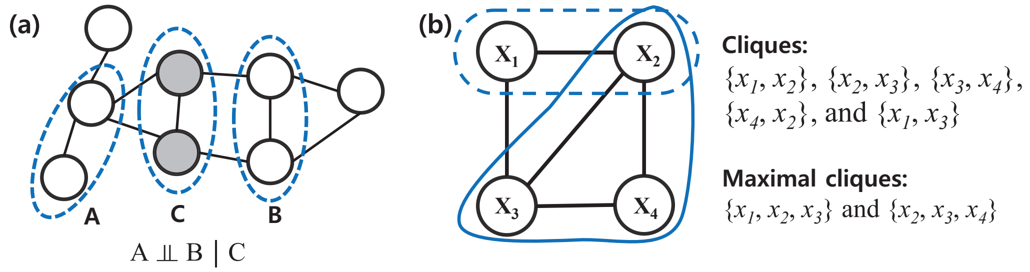


Figure 2.2: (a) An example of an undirected graph in which every path from any node in set A to any node in set B passes through at least one node in set C. Consequently the conditional independence property $A \perp\!\!\!\perp B|C$ holds for any probability distribution described by this graph. (b) A four-node undirected graph showing a clique (straight line) and a maximal clique (dotted line).

These concepts are illustrated by the undirected graph over four variables shown in Figure 2.2 (b). Let us denote a clique by C and the set of variables in that clique by $\mathbf{x}_C$. Then the joint distribution is written as a product of *potential function* $\psi_C(\mathbf{x}_C)$ over the maximal cliques of the graph

$$P(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathbf{x}_C), \tag{2.3}$$

where the quantity Z, called the partition function, is a normalization constant and is given by

$$Z = \sum_{\mathrm{x}}\prod_{C}\psi_C(\mathbf{x}_C) \tag{2.4}$$

which ensures that the distribution $p(x)$ given by (2.4) is correctly normalized.

In this manner, MRFs can represent circular dependencies between variables and it is useful in the cases in which direction of influence has no meaning, for example when variables represent pixels in image or atom in a protein molecule. As such, MRFs have seen wide application in many areas, including computer vision (Li, 1995, 2009; Wang et al., 2013), and bioinformatics (Demirkaya et al., 2005; Wei and Li, 2007; Chen et al., 2011).

2.2.3 Hidden Markov Models

A hidden Markov model (HMM) (Rabiner and Juang, 1986; Eddy, 1996; Bishop and Nasrabadi, 2006; Kollar and Friedman, 2009; Murphy, 2012) can be viewed as a specific instance of the state space model of Figure 2.3 in which the latent variables are discrete. The HMM models a sequence of observations $X = \{x_t\}_{t=1}^{T}$ by assuming that there is an underlying sequence of states $Y = \{y_t\}_{t=1}^{T}$ drawn from a finite state set S. To model the joint distribution $p(\mathbf{y}, \mathbf{x})$ tractably, an HMM makes two independence assumptions. First, it assumes that each state depends only on its immediate predecessor, that is, each state yt is independent of all its ancestors $y_1, y_2, ..., y_{t-2}$ given its previous state y_{t-1}. Second, an HMM assumes that each observation variable x_t depends only on the current state y_t. With these assumptions, we can specify an HMM using three probability distributions: first, the distribution $p(y_1)$ over initial states; second, the transition distribution $p(y_t|y_{t-1})$; and finally, the observation distribution $p(x_t|y_t)$. That is, the joint probability of a state sequence $\mathbf{y}$ and an observation sequence $\mathbf{x}$ factorizes as

$$p(\mathrm{y}, \mathrm{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t), \tag{2.5}$$

where to simplify notation, we write the initial state distribution $p(y_1)$ as $p(y_1|y_0)$.

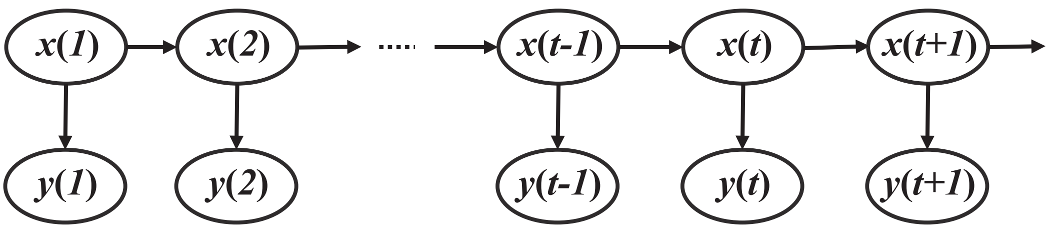


Figure 2.3: Graphical structure of hidden Markov model. We can represent sequential data using a Markov chain of latent variables, with each observation conditioned on the state of the corresponding latent variable.

The HMM is widely used in speech recognition (Rabiner, 1989), natural language modeling (Manning and Schütze, 1999), and for the analysis of biological data (Eddy, 1998; Krogh et al., 2001; Wheeler et al., 2013; Bonneville and Jin, 2013).

2.3 Higher-order Graphical Models for Biological Problems

2.3.1 Higher-Order Models

We generally assume pairwise relationships among the objects of our interest in machine learning problem setting. An object set endowed with pairwise relations can be naturally described as a graph, in which the vertices represent the objects, and any two vertices that have some kind of relationship are joined together by an edge. However, in many real-world problems, relationships among the objects of our interest are more higher-order than pairwise, and thus representing a set of their complex relationships as general undirected or directed graphs is not complete.

A higher-order model uses higher-order units as features. While linear models are difficult to reflect high order dependency embodied in the data, higher-order models can represent higher-order relationships, thus fitting the complex solution spaces including nonlinearity.

A higher-order unit can be defined to a feature represented with patterns or function values derived from raw attributes of given data (Roddick et al., 2008; Lehar et al., 2008). When A is a subset of $\mathbf{X}$=$\{x_1,\cdots,x_m\}$, the set of attributes of the given data, formally, a feature **f** is defined as follows:

$$f = g(A) \tag{2.6}$$

where g denotes an arbitrary function. When f is an identity function, **f** denotes a raw attribute. Also, f is a linear feature with weighted summation of the elements of A in case of $g = \sum_{x_i \in A} w_i x_i$. On the other hands, **f** becomes a higher-order feature when g is a function with two or more order such as sin, exp, and $\prod_{x_i \in A} x_i$.

In this dissertation, we use an individual represented with a conjunction of attribute values of data, and a population of them as a higher-order unit and a higher-order model, respectively. This conjunction-based individual representation enhances the interpretability of the models more compared with units based on numerical functions. Also, the individual and the population in our study can be represented with a hyperedge and a hypergraph.

2.3.2 Hypergraphs

A hypergraph (Berge, 1989; Gallo et al., 1993; Zhou et al., 2007) is a generalized graph for representing complex interactions. In the hypergraph construct, the edge in a conventional graph (which connects two vertices) is generalized to the hyperedge, which connects more than two vertices concurrently. A hyperedge is weighted by the strength of the higher-order dependency among its elements.

Unlike conventional graphs, hypergraphs are suitable for explicitly representing higher-order relationships among many features.

Formally, a hypergraph H is formulated as a triple $H = \{V, E, W\}$, where V, E, and W denote the sets of vertices v, hyperedges e, and hyperedge weights $w(e)$, respectively. A hyperedge of weight $w(e)$ is represented as a subset of V. Let $d(v)$ and $\delta(e)$ denote the degree of a vertex v and the degree of a hyperedge e, respectively. Each degree is then defined as follows:

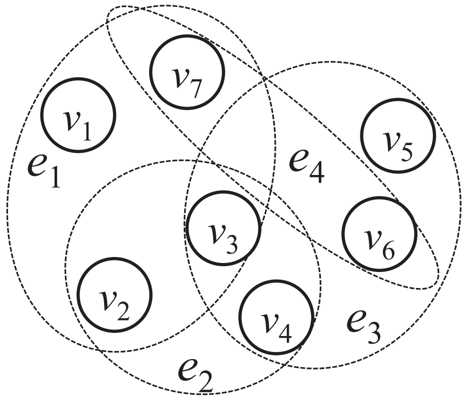
$$d(v) = \sum\nolimits_{e \in E} w(e)h(v, e) \tag{2.7}$$

and

$$\delta(e) = |e| \tag{2.8}$$

where |e| is the cardinality (number of vertices) of e and $h(v, e)$ is an indicator function that returns 1 if v is an element of e and 0 otherwise. A hyperedge with degree k is called a k-hyperedge and a hypergraph consisting solely of k-hyperedges is a k-hypergraph. Figure 2.4 shows an example of a hypergraph. A high-degree vertex can be regarded as a hub of the hypergraph structure, which may signify an informative feature for classifying the given data. Moreover, a hyperedge with higher degree embodies more specific information, whereas one with lower degree characterizes more general patterns. Because they naturally represent higher-order interactions, hypergraphs have become a popular choice for solving a range of problems (Hu et al., 2008; Klamt et al., 2009; Bu et al., 2010).

According to its applications, vertices and hyperedges denote different objects. In Zhou's study (Zhou et al., 2007), a vertex is a data instance and a hyperedge denotes a set of vertices with identical attribute values. In Kok's study to build Markov logic networks from relational database, on the other hands, a vertex represents



Hyperedges

Hyperedges	Vertices	$\delta(e)$	$w(e)$
e_1	v_1, v_2, v_3, v_7	4	2
e_2	v_2, v_3, v_4	3	1
e_3	v_3, v_4, v_5, v_6	4	3
e_4	v_4, v_7	2	4

Degree of vertices

$d(v_1) = 2$	$d(v_2) = 3$	$d(v_3) = 6$	$d(v_4) = 8$
$d(v_5) = 3$	$d(v_6) = 3$	$d(v_7) = 6$	-

Figure 2.4: An example of hyperedges in a hypergraph

a discrete data attribute (variable) and a hyperedge means logical relation among them (Kok and Domingos, 2009). Same to Kok's representation, a vertex means discrete variable value and a hyperedge represents an arbitrary combination of vertices in Zhang's models (Zhang, 2008).

The understanding of complex biological systems is a fundamental issue in computational biology. In particular, when analyzing topological properties of biological networks, one often tends to substitute the term "network" for "graph", or uses both terms interchangeably. From a mathematical perspective, this is not fully correct, because many functional relationships in biological networks are more complicated than what can be represented in graphs.

As mentioned above, graphs are combinatorial models for representing relationships (edges) between certain objects (vertices or nodes). In biology, the vertices typically illustrate genes, transcription factors, proteins, metabolites, or other biological components, whereas the edges represent functional relationships or inter-

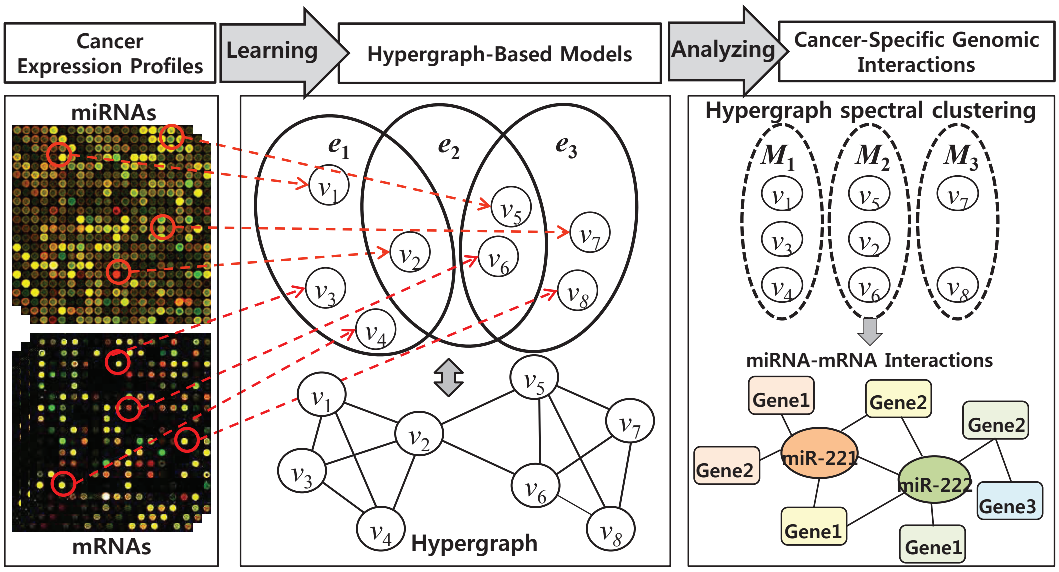


Figure 2.5: Modeling genomic interactions via hypergraph-based models

actions between the vertices such as “binds to”, “regulates to”, or “is converted to”. A key property of graphs is that every edge connects two vertices. However, many biological processes including cancer are characterized by more than two participating cooperators and are thus not bilateral. Hence, multilateral relations are not compatible with general graph edges. In addition, transformation to a graph representation is usually possible but may imply a loss of information that can lead to wrong interpretations subsequently.

Hypergraphs provide a framework that helps to overcome such conceptual limitations. As the name indicates, a hypergraph is a generalized graph by allowing edges to connect more than two vertices, which may facilitate a more precise representation of higher-order interactions in biological processes. Thus, hypergraph-based models are suitable for representing a knowledge network to investigate complex biological phenomena and they have been successfully used for diverse biological problems (Ha et al., 2007; Kim et al., 2007; Tian et al., 2008; Klamt et al., 2009; Kim et al., 2010). Figure 2.5 shows an example of a hypergraph-based models for modeling cancer-specific genomic interactions from cancer expression profiles.

Chapter 3

Hypergraph Classifiers for Identifying Prognostic Modules in Cancer

3.1 Overview

Predicting the clinical outcomes of cancer patients is a challenging task in biomedicine. A personalized and refined therapy based on predicting prognostic outcomes of cancer patients has been actively sought in the past decade. Accurate prognostic prediction requires higher-order representations of complex dependencies among genetic factors. However, identifying the co-regulatory roles and functional effects of genetic interactions on cancer prognosis is hindered by the complexity of the interactions.

In this chapter, we introduce a new population-based model that uses an evolutionary learning method to predict clinical outcomes of cancer patients (Figure 3.1) (Kim et al., 2013a). The model handles complex genomic interactions by means

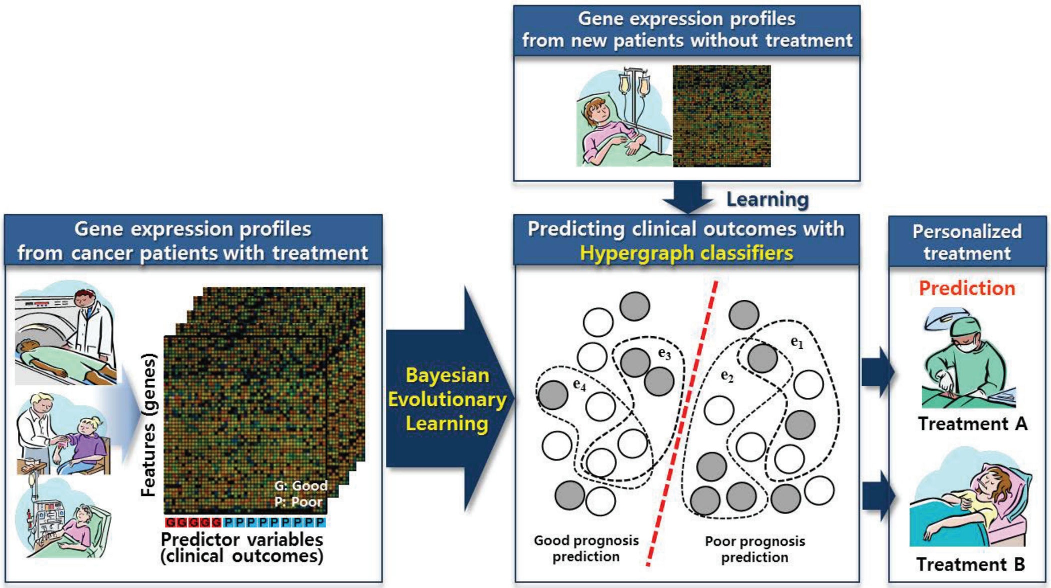

Gene expression profiles from new patients without treatment
Learning
Gene expression profiles from cancer patients with treatment
Features (genes)
G: Good
P: Poor
GGGGGPPPPPPPPPP
Predictor variables (clinical outcomes)
Bayesian Evolutionary Learning
Predicting clinical outcomes with Hypergraph classifiers
e1
e2
e3
e4
Good prognosis prediction
Poor prognosis prediction
Personalized treatment
Prediction
Treatment A
Treatment B

Figure 3.1: Overview of the hypergraph classifier based on Bayesian evolutionary learning for predicting cancer clinical outcomes from cancer genomic data

of a flexible hypergraph structure comprising a large population of hyperedges, representing the multi-variable combinations corresponding to all potential genes or markers. Each hyperedge is weighted by its discriminative ability to predict prognostic outcomes. Thus, each hyperedge potentially behaves as a prognostic module influencing the cancer clinical outcomes.

The model learning involves the search of a high-dimensional space reflecting the higher-order relationships between factors. To learn the model from a dataset comprising several tens of thousands of genetic variables, an evolutionary method based on sequential Bayesian sampling scheme is applied (Ha et al., 2013). The proposed Bayesian evolutionary algorithm is designed upon a standard evolutionary computation framework. Variation, evaluation, and selection are repeated as a sequential Bayesian sampling process, where the posterior distribution is recursively calculated from the prior distribution by estimating the likelihood from fitness measurements. Using this Bayesian formulation of evolutionary computation, the model can determine the problem-specific bias as a guideline for efficient search of a huge combinatorial feature space. This study adopts an information theoretic co-regulatory measure called mutual information, and the model complexity for the distribution. The information theoretic measure enhances the efficiency of the evolutionary search, while the complexity retains a compact model size by controlling the parsimony.

The proposed model is evaluated on MAQC-II breast cancer and multiple myeloma gene expression data (Shi et al., 2010). The proposed model demonstrates high classification performance for predicting prognosis in patients, and can identify higher-order prognostic biomarkers of cancer clinical outcomes. Moreover, our model directly identifies potential modules of informative genes that characterize prognosis and recurrence risk in cancer.

3.2 Analyzing Gene Modules for Cancer Prognosis Prediction

Prognostic prediction is an important task in clinical medicine. Estimating the clinical outcomes of patients and the potential effects of treatment is crucial. A refined treatment based on likely clinical outcomes is especially necessary in oncology, because cancer progression varies between patients. By accurately estimating the clinical response to treatment, clinicians can personalize and hence provide an improved therapy for a patient.

Gene expression profiling has been widely used to identify tumor heterogeneity, and has led to the discovery of molecular signatures of potential prognostic and therapeutic interest (Simon, 2003; Fan et al., 2010; Goodison et al., 2010). As such, it is recognized as a powerful source for improving prognostic assessment and treatment selection in cancer medicine. Moreover, cancer prognosis is associated with combinatorial and modular regulation by multiple genetic factors. Thus, for more precise prediction of cancer clinical outcomes, the higher-order relationships among genetic factors must be deduced from gene expression profiles. However, the complexity of gene interactions renders this task extremely challenging.

Predictive methods, which classify patient outcomes on a genome-wide scale from high-throughput biological data, have flourished in recent decades. Many studies have adopted computational approaches, such as machine learning-based models (Veer et al., 2002; Street et al., 1995; Koziol et al., 2009; Sun et al., 2011; Verduijn et al., 2007; Gevaert et al., 2006; Han et al., 2011; Berchuck et al., 2005; Kim et al., 2012a) and statistical methods (Braitman and Davidoff, 1996; Huang et al., 2003; Boulesteix et al., 2008; Matsui et al., 2007), to predict prognosis from cancer genomic data. However, few of the existing approaches address the higher-order interactions between genes involved in cancer prognosis.

Predicting outcomes from higher-order gene relationships requires searching of an exponential search space consisting of tens of thousands of genes. Such a huge combinatorial feature space cannot be exhaustively searched using a gradient method, and is instead undertaken by various feature selection methods (Saeys et al., 2007). Typically, these approaches reduce the problem space by individually evaluating each gene, assuming independence between features. However, such restrictions may not capture the important genes involved in higher-order relationships underlying pathological processes.

3.3 Hypergraph Classifiers for Identifying Cancer Gene Modules

3.3.1 Hypergraph Classifiers

The proposed population-based model uses hypergraph structures composed of a large collection of hyperedges playing the role of a weak classifier. These hyperedge ensembles are called hypergraph classifiers. The unlabeled data can be predicted by assembling this population of many weak classifiers.

We assume that in the n-th data instance, a set of class labels denoting clinical outcome, Y, and a hypergraph, H, are given. The y value whose weighted sum of hyperedges corresponding to the genetic variables in $\mathbf{x}^{(n)}$ is the largest among the elements of Y is called the class label of $\mathbf{x}^{(n)}$, denoted $\hat{y}^{(n)}$. Specifically, the class label is determined as follows:

1. Calculate c_y, the sum of weights for $y \in$Y over all hyperedges in E:

$$c_y = \sum_{i=1}^{|E|} \left\{ w(e_i) f(\mathbf{x}^{(n)}, e_i) \varphi(y^{(n)}, y_i) \right\} \tag{3.1}$$

   where $|E|$ denotes the hyperedge set and $w(e_i)$ is the weight of e_i.

2. Predict the class label of $\mathbf{x}^{(n)}$, $\hat{y}^{(n)}$, as the y value with the highest total weight:

$$\hat{y}^{(n)} = \arg\max_{y \in Y} c_y \tag{3.2}$$

In Equations (3.1) and (3.2) above, $f(\mathbf{x}^{(n)}, e_i)$ and $\varphi(y^{(n)}, y_i)$ denote the matching and indicator functions, which return 1 if e_i matches $\mathbf{x}^{(n)}$ and if $y^{(n)} = y_i$, respectively. These functions are defined as Equations (3.14) and (3.15) in the next subsection. This classification process is similar to an learning classifier system (LCS) (Holland, 1980), in which each classifier participates in classifying the unlabeled data as a significant condition-action rule. However, the proposed hypergraph classifier focuses on the model structure (the entire connected ensemble of hyperedges), rather than on each hyperedge. The hyperedges composing the population exert the main influence on the classification performance. In the next subsection, we explain how the population is generated and how the model is learned by an evolutionary method.

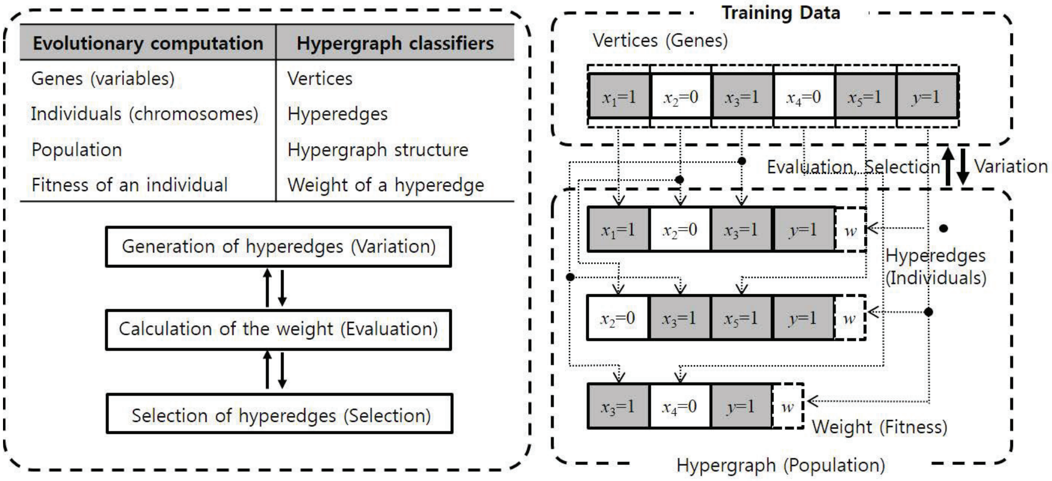
3.3.2 Bayesian Evolutionary Algorithm

The Bayesian evolutionary algorithm implements an evolutionary learning method based on sequential Bayesian sampling. A standard evolutionary computation process that iterates the generation of individuals (variation), calculation of the fitness (evaluation), and selection of individuals (selection) is implemented with the Bayesian sampling framework where the posterior distribution is recursively computed by estimating the likelihood from the prior distribution. Figure 3.2 presents the terms of hypergraph classifiers and their corresponding terms in standard evolutionary computation schema. A naive evolutionary method may be inefficient when the problem involves the searching of vast and complex solution spaces. However, Bayesian evolutionary algorithm can efficiently search the space by introducing problem-specific knowledge to the prior distribution.

Let H_t be a population at the t-th generation. For a dataset $D = (X, Y)$, where $X = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ and $Y = \{y^{(n)}\}_{n=1}^{N}$ are given, Bayes' rule specifies the posterior distribution of H_t as the conditional probability:

$$p(H_t|X, Y) = \frac{p(Y|X, H_t)p(H_t|X)}{p(Y|X)} \tag{3.3}$$

where $p(Y|X, H_t)$ and $p(H_t|X)$ denote the likelihood and the prior, respectively. In Equation (3.3), $p(Y|X)$ is a normalizing constant because it is independent of H_t.



Evolutionary computation	Hypergraph classifiers
Genes (variables)	Vertices
Individuals (chromosomes)	Hyperedges
Population	Hypergraph structure
Fitness of an individual	Weight of a hyperedge

Figure 3.2: Hypergraph classifiers in standard evolutionary computation

Thus, the posteriori distribution is proportional to the product of the likelihood and the prior:

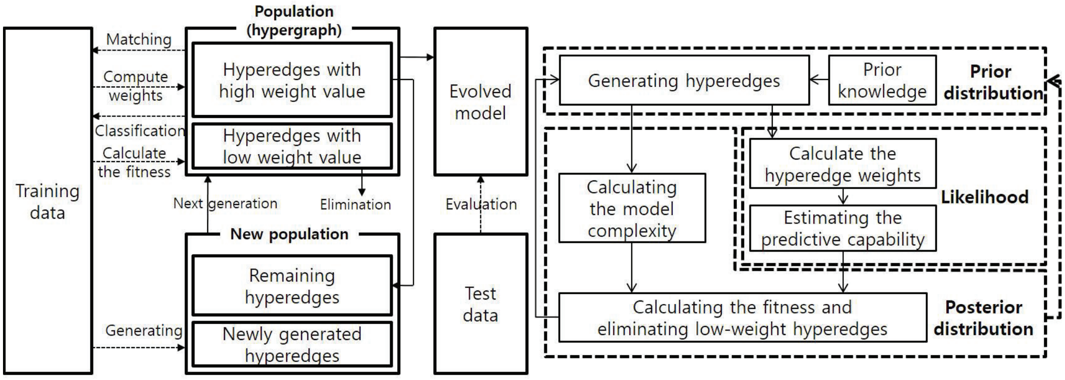
$$p(H_t|X,Y) \propto p(Y|X,H_t)p(H_t|X) \tag{3.4}$$

The aim of the evolutionary process is to maximize the model fitness F_t, defined as the logarithm of the posterior:

$$F_t = \log p(Y|X,H_t) + \log p(H_t|X) \ , \tag{3.5}$$

$$H^* = \arg\max_{H_t} F_t \tag{3.6}$$

Finally, the evolution of hypergraph classifiers is regarded as the maximum a posteriori (MAP) process in the Bayesian learning. Figure 3.3 describes the evolving process of hypergraph classifiers learned by the Bayesian approach.



(a) Evolving flow of hypergraph classifiers

(b) Bayesian view of evolving hypergraph classifiers

Figure 3.3: Flowchart of Bayesian evolutionary learning of hypergraph classifiers

3.3.3 Bayesian Evolutionary Learning for Hypergraph Classifiers

The model fitness is computed from the prior and the likelihood. The empirical prior distribution $p(H_t|X)$ can be defined by prior knowledge of the problem, which enhances the efficiency of the evolutionary search. In this study, the prior includes two factors. One is mutual information (MI) between each variable and the class label, specified as the relationships between the data rather than a uniform distribution. MI is an information-theoretic measure that specifies the degree of conditional independence between two random variables. Here, it is used as a co-regulatory measuring criterion for efficiently selecting genes for hyperedge generation. The other factor is the model complexity. The prior is defined to prefer the most parsimonious, or smallest, model. This prior not only ensures that genetic variables relevant to prognostic outcomes are more frequently included in the model, but also retains the model compact. Therefore, the current empirical prior for generating hyperedges is calculated from the MIs and the previous posterior, $p(H_{t-1}|Y, X)$

reflecting the model complexity $|H_{t-1}|$:

$$p(H_t|X) \propto p(H_{t-1}|Y, X) \tag{3.7}$$

$$p(H_t|X) \propto \frac{1}{|H_{t-1}|} \prod_{e \in E_{new}} P(e) \approx \frac{1}{|H_{t-1}|} \prod_{e \in E_{new}} \prod_{x_i \in e} P_I(x_i) \tag{3.8}$$

s.t.

$$P_I(X_i) = \frac{I(X_i;Y)^\eta}{\sum\limits_{j=1}^{|X|} \{I(X_i;Y)^\eta\}}, \ |H_t| = \sum\nolimits_{e \in E_t} \delta(e), \text{ and } E_{new} = E_t - E_{t-1},$$

where E_t is the hyperedge set of H_t, and $P(e)$ denotes the probability with which a hyperedge e is generated. $P_I(X_i)$ indicates the probability of selecting the i-th genetic factor X_i, which depends on the MI between X_i and the class label Y, $I(X_i;Y)$. The nonnegative constant η regulates the influence of MIs on the gene selection. The prior distribution influences hyperedge construction in every generation. Specifically, a hyperedge is generated as follows:

1. Select the data instance from which to subsample a hyperedge.

2. Probabilistically determine the degree of the hyperedge within a predefined range:

$$P(\delta(e) = K) = \frac{|E_{t-1}^K| + \varepsilon}{\sum\limits_{k=K_{\min}}^{K_{\max}} \left(|E_{t-1}^k| + \varepsilon\right)} \tag{3.9}$$

where E_{t-1}^k denotes a set of k-hyperedges at generation t-1 and ε is a smoothing constant.

3. Probabilistically select the variables based on $P_I(X_i)$.
4. Construct a hyperedge from a set of variable values and the class label of the selected data instance.
5. Add the generated hyperedges to the population.

Hyperedge generation in our model differs from that of LCS, where each classifier is generated by genetic operations such as crossover and mutation. Our model can efficiently search a high-dimensional space without a heavy computational cost, because it guarantees that a pattern in a hyperedge always exists in the training data.

The likelihood is defined to represent the discriminative capability of the model. To achieve this, we assume that the capability grows by increasing the difference of the weighted sum between the correctly and incorrectly matched hyperedges for all training data. A hyperedge is said to be correctly matched if it matches a given data instance and the label of the hyperedge equals that of the instance. On the other hand, an incorrectly matched hyperedge is matched to an instance with a different class label than itself. Since the instances are independent, the likelihood is estimated as the product of the empirical likelihoods on the given data:

$$p(Y|X, H_t) = \prod_{n=1}^{N} p(y^{(n)}|x^{(n)}, H_t) \tag{3.10}$$

and the empirical likelihood is defined by:

$$p\left(y^{(n)}|x^{(n)}, H_t\right) \equiv \frac{\sum_{i=1}^{|E_t|} w(e_i)\left\{f_i^{(n)} \cdot \varphi_i^{(n)} - f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)}\right)\right\}}{\sum_{i=1}^{|E_t|} w(e_i)} \tag{3.11}$$

where $w(e_i)$ denotes the weight of the i-th hyperedge. Thus, we have

$$p(Y|X,H_t) = \prod_{n=1}^{N} \left[\frac{\sum_{i=1}^{|E_t|} w(e_i) \left\{ f_i^{(n)} \cdot \left(2\varphi_i^{(n)} - 1 \right) \right\}}{\sum_{i=1}^{|E_t|} w(e_i)} \right] \tag{3.12}$$

$$\text{s.t. } f_i^{(n)} = f(x^{(n)}, e_i) \text{and} \varphi_i^{(n)} = \varphi(y^{(n)}, y_i) \tag{3.13}$$

with the matching and indicator functions respectively defined as follows:

$$f_i^{(n)} = f(x^{(n)}, e_i) = \begin{cases} 1, \text{if} \exp\left\{ c(x^{(n)}, e_i) - \left| e_i - \{y_i\} \right| \right\} > \theta \\ 0, \text{otherwise} \end{cases}, \tag{3.14}$$

$$\varphi_i^{(n)} = \varphi(y^{(n)}, y_i) = \begin{cases} 1, \text{if} y^{(n)} = y_i \\ 0, \text{otherwise} \end{cases}, \tag{3.15}$$

where $c(x^{(n)}, e_i)$ is the matching number, defined as the number of hyperedge variables that equal their corresponding variables in $\mathbf{x}(n)$. The matching threshold θ smoothes and enhances robustness against data noise by allowing partial matching. Also, $f_i^{(n)} \cdot \varphi_i^{(n)}$ and $f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)}\right)$ equal 1 for a correctly and incorrectly matched hyperedge, respectively, and 0 otherwise. The weight of a hyperedge is a function of correctly and incorrectly matched cases:

$$\begin{aligned} w(e_i) &= \left| y_i \right|^{\beta} \left\{ \alpha \sum_{n=1}^{N} f_i^{(n)} \cdot \varphi_i^{(n)} - (1-\alpha) \sum_{n=1}^{N} f_i^{(n)} \cdot \left(1 - \varphi_i^{(n)} \right) \right\} && (3.16) \\ &= \left| y_i \right|^{\beta} \left\{ \sum_{n=1}^{N} f_i^{(n)} \cdot \varphi_i^{(n)} - (1-\alpha) \sum_{n=1}^{N} f_i^{(n)} \right\}, && (3.17) \end{aligned}$$

where α is a constant for preferring more correct or less incorrect predictions. For data whose class labels are imbalanced, a quantity $|y_i|$, denoting the number of data with class label y_i, and a negative constant β, are introduced into the weight function. If $w(e)$ is negative, it is reset to zero to prevent the construction of a negatively weighted graph. The model fitness is then reformulated from (3.5) using the defined prior (3.8) and the estimated likelihood (3.13):

$$F_t = \log p(Y|X, H_t) + \log p(H_t|X)$$

$$\approx \sum_{n=1}^{N} \log \frac{\sum_{i=1}^{|E_t|} \left\{ w(e_i) f_i^{(n)} \left(2\varphi_i^{(n)} - 1 \right) \right\}}{\sum_{i=1}^{|E_t|} w(e_i)} + \lambda \, |H_t| \, + \zeta \sum_{e \in E_t} \log \sum_{x_i \in e} P_I(x_i) \tag{3.18}$$

where λ and ζ denote a negative constant for regularizing the model size and a positive value for regulating the selection power of the variables in the prior, respectively. To increase the fitness, hyperedges with high weight survive at every generation; in addition, a hyperedge is generated from variables with large $P_I(x)$, and the proportion of lower-degree hyperedges is increased.

As the population changes, the hypergraph structure evolves by removing hyperedges with relatively low weight and replacing them with new hyperedges at every generation. To prevent the removal of highly discriminating hyperedges, the number of replaced hyperedges decreases to a specific value as the iterations proceed. The number of replacements at the t-th generation is adaptively determined:

$$R_t = \frac{R_{max} - R_{min}}{\exp\left(t/\kappa\right)} + R_{min} \tag{3.19}$$

where t is the iteration number of the learning process, and R_{max} and R_{min} denote the maximum and minimum boundary values of R_t, respectively. κ is a constant that moderates the speed at which the system proceeds from R_{max} to R_{min}.

3.4 Predicting Cancer Clinical Outcomes Based on Gene Modules

3.4.1 Data and Experimental Settings

The gene expression data have been widely used in various applications. They include diagnosis, early detection, monitoring of disease progression, risk assessment, prognosis, complex medical product characterization and prediction of response to treatment. For this reason, many classification models for microarray data have been proposed for being applied to the biological and medical fields. Herein, the published benchmarking studies on classifiers for microarray data have split data into two sets: a dataset used for training and the other set for validation, with randomness. This design assumes that the training and validation sets are produced by unbiased sampling of a large and homogeneous population of samples. However, specimens in clinical studies are usually accrued over years and there may be a shift in the participating patient population and also in the methods used to assign disease status owing to the change of practice standards.

The MicroArray Quality Control (MAQC)-II project (Shi et al., 2010) was designed to evaluate these sources of bias in study design by constructing training and validation sets at different times, swapping the test and training sets and also using data from diverse preclinical and clinical scenarios. The goals of MAQC-II were to survey approaches in genomic model development in an attempt to understand the sources of variability in prediction performance and to assess the influences of endpoint signal strength in data. Thus, the use of the MAQC-II datasets can enhance our capability to more accurately predict the clinically relevant cancer prognosis.

The proposed model is evaluated on MAQC-II gene expression data of human breast cancer and multiple myeloma. The breast cancer dataset consisting of 12,993 genes is used to predict pathological complete response (pCR) to preoperative chemotherapy. It was originally divided into two sets: a 130-sample training set consisting of 33 positives and 97 negatives, and a 100-sample test set consisting of 15 positives and 85 negatives. The multiple myeloma dataset consisting of 20,638 genes is used to predict the overall survival (OS) 730 days post-treatment. The original 340-sample training set consisted of 51 positives and 289 negatives, while the 214-sample test set comprised 27 positives and 187 negatives. During preprocessing, sample-wise and feature-wise normalization was conducted, and the variable data values were converted into three-level discretized values {-1, 0, 1} based on z-scores.

The experimental parameter settings are listed in Table 3.1. The parameters are determined as the values yielding optimal performance after empirical experiments. Although a hypergraph classifier has many parameters, most of them can be used as default values independent on problems. Main parameters determined according to problems are initial population size and individual length. Too small initial population causes the discriminative capability of the model to decrease due to the lack of the information for classification. Too large population size leads too heavy computational cost. Therefore, the appropriate range of initial population size is from five to one hundred. Individual length influences the discriminative ability and the probability matching data of a hyperedge. The minimum value of the length is usually set to three and the maximum value does not usually exceed ten. The proper ranges of the parameter values are presented in Table 3.1. To investigate the effect of the Bayesian evolutionary learning method on classification performance, experiments were conducted under various parameter conditions on the model prior.

Table 3.1: Parameter settings of the proposed model used in experiments

Terms	Description	BC dataset	MM dataset
Initial Pop. size	Number of hyperedges	5 **x** $\|D^{tr}\|$	1 **x** $\|D^{tr}\|$
Individual length	Degree of a hyperedge	Min:3, Max:6	Min:3, Max:6
λ	Regularization of the model size	0.001	0.001
ζ	Ratio of MI in fitness value	0.01	0.01
η	Reflecting MI values	1	1
α	Weighting the positive matching function value	0.1	0.1
β	Constant for imbalanced data	1	1
θ_L, θ_C	Matching threshold for learning and classification	0.9, 0.9	0.9, 0.9
R_{max}, R_{min}	Max. and Min. amounts of removed hyperedges	R_{max}: 0.5 **x** $\|E_t\|$ R_{min}: 0.1 **x** $\|E_t\|$	R_{max}: 0.5 **x** $\|E_t\|$ R_{min}: 0.1 **x** $\|E_t\|$
Iteration Number	Condition for terminating the evolution	30	20

BC and MM denote breast cancer and multiple myeloma, respectively.

3.4.2 Prediction Performance

Classification performance was evaluated using six standard classification models: Naive Bayes classifier, random forest (the number of trees = 10), AdaBoost with J48, and support vector machine (SVM) with sequential minimal optimization (SMO) and the second polynomial kernel implemented in Weka (Hall et al., 2009). A variant of learning classifier system (LCS), sUpervised Classifier System (UCS), were also used (Edakunni et al., 2009). We used default values of Weka as the parameters not explained of the other models. In LCS, the pop-size and the iteration number are 1000 and 500, respectively. Because of the large number of variables, probability of the wild card is set to 0.9997. The classification performance of each model was evaluated using the original validation datasets from the MAQC-II project. The results of the evolutionary learning-based models (our model and LCS) were averaged over 10 runs on each test dataset. Prediction performance was based on four measures; sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC), defined below:

$$\begin{aligned}
\text{Sensitivity} &= \tfrac{TP}{TP+FN}, \\
\text{Specificity} &= \tfrac{TN}{FP+TN}, \\
\text{Accuracy} &= \tfrac{TP+TN}{TP+FN+FP+TN}, \\
\text{MCC} &= \tfrac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}
\end{aligned} \tag{3.20}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. In particular, MCC is informative when the ratio of two classes in a dataset is highly skewed. Consequently, MCC has become a popular reference performance measure in bioinformatics, biomedical informatics, and other fields involving unbalanced datasets. MCC values range from +1 to −1, where +1 indicates a perfect prediction, 0 is essentially random prediction, and −1 is the asymptote of extreme misclassification.

Table 3.2 and 3.3 present the performance of the proposed model compared with other models. As revealed by the adjusted *p*-values, the accuracy of hypergraph

Table 3.2: Comparison of classification performance on breast cancer test dataset

Models	Sensitivity	Specificity	Accuracy	MCC
HC	0.45	0.90	0.84	0.37
2-HC	0 (7.8e-3)	1 (3.9e-3)	0.85 (1.9e-1)	N/A (-)
HC (no MI)	0.04 (7.8e-3)	0.97 (3.9e-3)	0.83 (1.9e-1)	0.11 (7.8e-3)
NB	0.73 (7.8e-3)	0.6 (3.9e-3)	0.62 (9.7e-3)	0.23 (1.7e-2)
RF	0.2 (1.1e-2)	0.81 (3.9e-3)	0.72 (9.7e-3)	0.01 (7.8e-3)
Ada	0.53 (4.5e-1)	0.81 (3.9e-3)	0.77 (9.7e-3)	0.28 (4.9e-2)
SVM	0.46 (4.5e-1)	0.89 (1.5e-1)	0.83 (1.9e-1)	0.35 (6.7e-1)
LCS	0 (7.8e-3)	1 (3.9e-3)	0.85 (1.9e-1)	N/A (-)

HC: Hypergraph Classifiers, 2-HCs: hyperedges of degree 2 (excluding the class vertex), HC (no MI): do not use MI as prior, NB: Naive Bayes, RF: Random Forest, Ada: AdaBoost (J48), SVM: Support Vector Machine, LCS: Learning Classifier System (UCS).

The performance of each model was evaluated using the original test (validation) datasets. All results are averaged over 10 runs on each test dataset. 'N/A' denotes no value and it occurs when TP+FP or FN+TN is zero because a model classifies all data as a certain class label. Values in the parenthesis denote adjusted p-values by multiple comparison correction (Bonferroni correction).

classifiers is similar to those of SVM and LCS, and superior to those of naive Bayes classifier, decision tree, random forest, and AdaBoost on both datasets. The adjusted p-values are calculated based on Wilcoxon signed-ranks test and multiple comparison correction with Bonferroni correction. Compared to existing models, the MCC obtained by our hypergraph model is especially improved on the multiple myeloma dataset with a significant adjusted p-value. Although LCS and SVM

Table 3.3: Comparison of classification performance on multiple myeloma test dataset

Models	Sensitivity	Specificity	Accuracy	MCC
HC	0.33	0.92	0.84	0.32
2-HC	0 (1.9e-3)	1 (5.8e-3)	0.87 (7.8e-3)	N/A (-)
HC (no MI)	0.05 (1.9e-3)	0.97 (5.8e-3)	0.86 (7.8e-3)	0.06 (1.9e-3)
NB	0.55 (1.9e-3)	0.69 (5.8e-3)	0.67 (7.8e-3)	0.17 (1.9e-3)
RF	0.03 (1.9e-3)	0.97 (5.8e-3)	0.85 (5.2e-1)	0.02 (1.9e-3)
Ada	0.22 (1.9e-3)	0.89 (1.4e-1)	0.82 (2.3e-2)	0.22 (1.9e-3)
SVM	0 (1.9e-3)	0.99 (5.8e-3)	0.86 (7.8e-3)	-0.02 (1.9e-3)
LCS	0 (1.9e-3)	1 (5.8e-3)	0.87 (7.8e-3)	N/A (-)

The results are obtained under the same condition as Table 3.2

demonstrate strong prediction accuracy, another measure is necessary for more precisely measuring the prediction capability in these problems, because the accuracy is distorted by severe imbalance of the classes in the datasets. Therefore, the proposed hypergraph classifiers more precisely predict clinical outcomes than existing models in terms of MCC and sensitivity. In addition, comparing the results of HC and 2-HC (a hypergraph classifier with degree-2 hyperedges), we observe that higher-order relationships are more important for accurately predicting cancer prognosis than pair-wise relationships. Moreover, we note that the model performance of HCs using MI as prior is improved by efficient searching of the huge combinatorial space.

Figure 3.4 plots the receiver operating characteristic (ROC) curves of the pro-

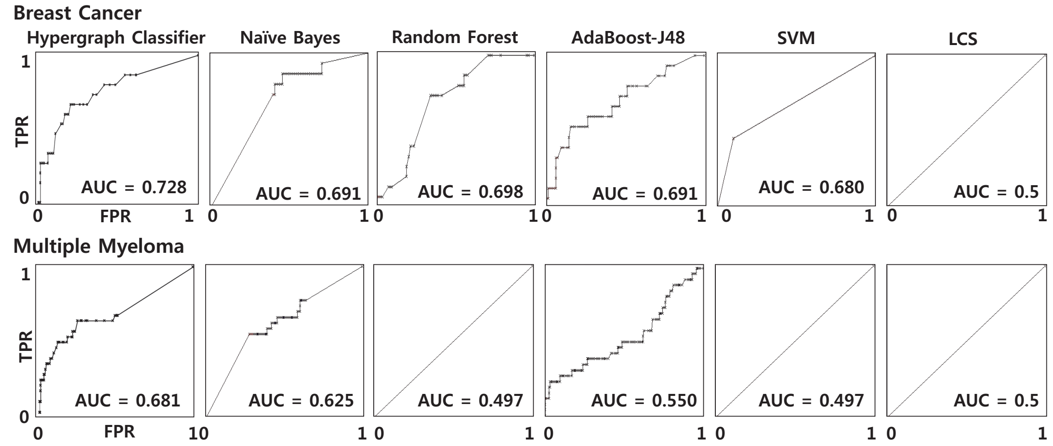


Figure 3.4: ROC curves with AUC of the proposed hypergraph classifier and other models on the test datasets of breast cancer (above) and multiple myeloma (below). TPR (true positive rate) and FPR (false positive rate) denote sensitivity and 1-specificity.

posed hypergraph classifier and other classification models on the test datasets of breast cancer and multiple myeloma, respectively. The areas under the ROC curves (AUCs) are calculated as a measure of predictive discrimination in the given test dataset in terms of specificity and sensitivity. An index of 0.5 presents no discrimination ability, whereas a value of 1 indicates perfect discrimination. Our model showed better classification performance than other models considering AUCs in Figure 3.4 and this result is consistent with that presented in Table 3.2 and 3.3. Interestingly, NB shows relatively high AUC compared to other measures in multiple myeloma and this is caused by the property that an AUC is large when the difference between sensitivity and specificity is small. From these results, we indicate that the hypergraph classifier is suitable model for classifying imbalanced data with

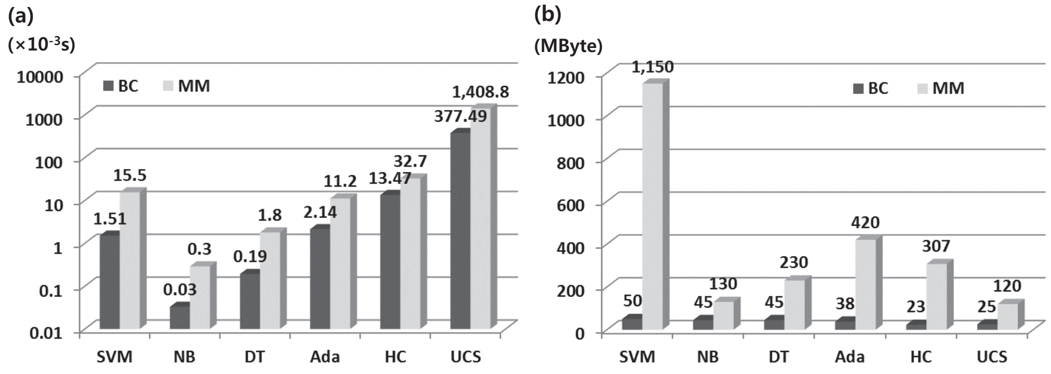
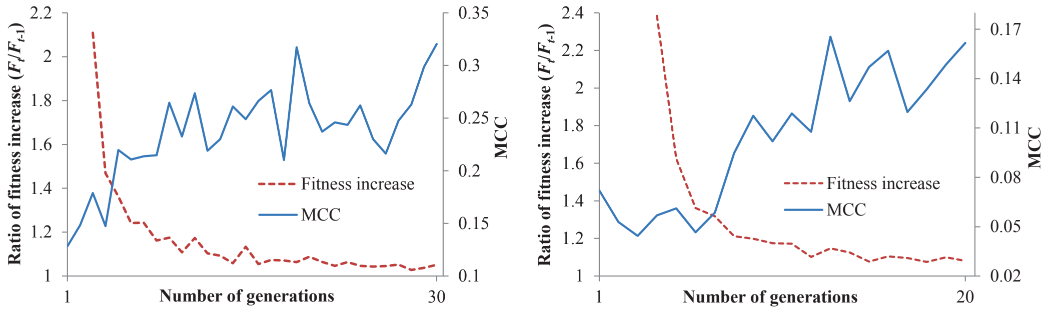


Figure 3.5: The proposed model of (a) time cost and (b) memory size in learning from breast cancer (BC) and multiple myeloma (MM) test data.

high dimensionality compared to other models.

The proposed hypergraph model belongs to a memory-based approach and the model complexity mainly depends on three terms such as the data size, the number of hyperedges, and the hyperedge degrees. Considering that hyperedge degrees can be considered as a constant, the time complexity is O(MN), where M and N denote the number of hyperedges and the data size. Moreover, the number of features increases the model complexity because the size of features usually influences the sufficient number of hyperedges due to the exponential increase of the model space. Figure 3.5 (a) and (b) show the time spent and the memory size used in learning from breast cancer and multiple myeloma dataset, respectively. Our model spends more time in learning compared to other machine learning methods and requires less time than learning classifier systems. The computational environment for the experiments involves Intel Xeon X5690 with 24 cores and 64 Gigabyte RAM based on Window 7 64bit.



(a) Breast cancer

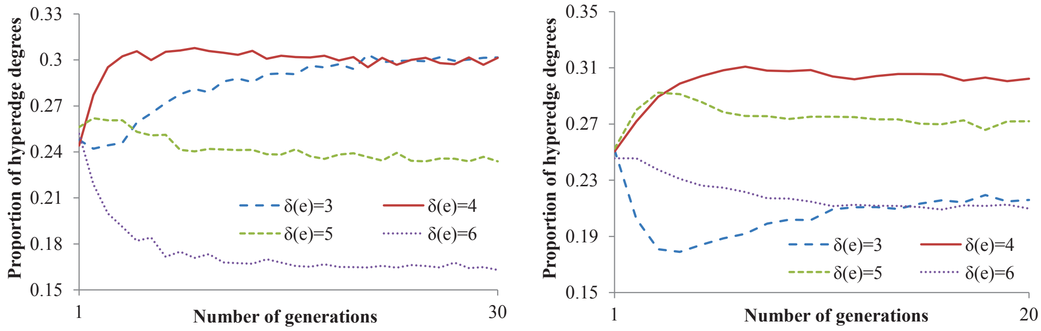
(b) Multiple myeloma

Figure 3.6: MCC fitness dynamics of the evolving hypergraph classifiers, evaluated on test datasets. The results are averages of 10 runs.

3.4.3 Model Analysis

We now present the changes of the proposed model as the Bayesian evolution proceeds. Figure 3.6 shows the dynamics of the MCCs and fitness values evaluated on the breast cancer and multiple myeloma datasets, respectively. Although the MCCs fluctuate, they increase overall as the learning proceeds. The fitness values increase toward their specified maximum. Thus, the defined fitness function reasonably indicates the discriminative capability of the model. In addition, the proposed model evolves into a predictive model that is competitive in terms of both accuracy and MCC despite the skewed class ratio of the data.

Next, we explored the evolution of the hypergraph classifier structure, by analyzing the composition of the hyperedges. The dynamics of hyperedge degree distribution are plotted in Figure 3.7. For both datasets, the proportion of lower-degree hyperedges (δ(e) = 3 and 4) increases as the number of generations increases, while the proportion of higher-order degree hyperedges (δ(e) > 5) decreases. Lower-

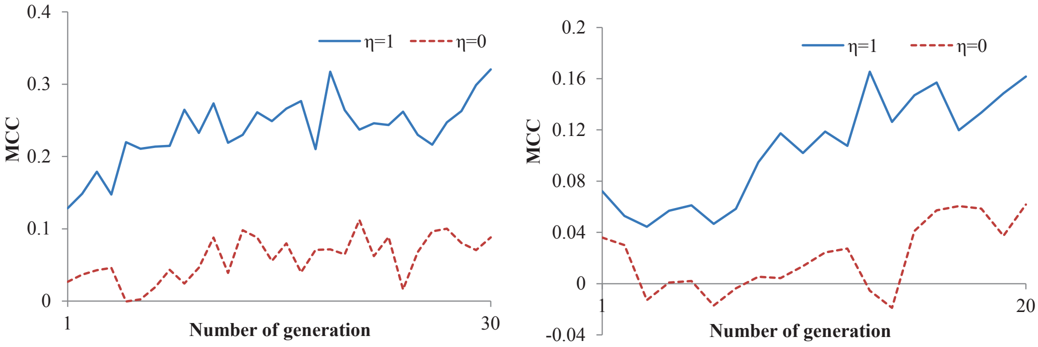


(a) Breast cancer (b) Multiple myeloma

Figure 3.7: Changes in the distribution of the degree of hyperedges in the evolving hypergraph classifiers. The y-axis denotes the proportion of k-hyperedges to $|E_t|$.

degree hyperedges are assigned a higher weight to reflect their higher probability of matching more training data. In Figure 3.7 (b), especially, 3-hyperedges steadily increase following a decrease in early generations. This initial decrease occurs because, although 3-hyperedges are more likely to match training data, they are also prone to incorrect matching. However, highly discriminative 3-hyperedges survive under the evolutionary learning and thus their proportion increases. Furthermore, higher-degree hyperedges with $\delta(\mathrm{e}) > 5$ are useful for class discrimination because their proportion never converges to zero. Higher-order hyperedges may be especially important for classifying data involving complex relationships between factors. According to Figure 3.7, the proportion of 5-hyperedges ($\delta(\mathrm{e}) = 5$) increases during the early stages of the evolution, and subsequently decreases. This pattern typifies evolutionary phenomena in nature, suggesting that 5-hyperedges play the role of intermediates in the evolutionary process.

Figure 3.8 shows how the learning performance of the model depends on MI



(a) Breast cancer

(b) Multiple myeloma

Figure 3.8: MCC dynamics of the hypergraph classifiers according to MI. $\eta = 0$ denotes that MI as prior was not used.

used as the prior. The effect of the prior on evolving hypergraph classifiers can be investigated by varying the parameter η. From (3.8), when $\eta = 0$, the model reduces to naive random search-based evolution. We observe that MI improves the efficiency of the learning and increases the performance of the model throughout the evolution.

3.4.4 Identification of Prognostic Gene Modules

Here, we analyze the structure of the hypergraph classifiers at the hyperedge level as the model is evolved. Table 3.4 and 3.5 list the genes with large $d(v)$ and the degree of vertices included in hypergraph classifiers learned from each dataset, together with their MI-rank. Genes with large $d(v)$ can be regarded as genes that significantly affect prediction. The threshold of $d(v)$ is defined as the $d(v)$ for which $p < 0.05$, determined by averaging $d(v)$ over all genes. As shown in Table 3.4 and 3.5, many genes with both high and low MI rank appear in the list of large $d(v)$. Those

Table 3.4: List of genes with high $d(v)$ in breast cancer data, identified by the learned model ($p < 0.05$)

♯ App.	Genes (MI-rank)
10	FERMT1(2), SNED1(3), PTGER3(5), HECA(9), MKI67(11), SOX11(12), JMJD6(14), NUCB2(16), FAM153A(19), GREB1(20), TMED7(21), TMEM48(22), KLHDC2(23), GATA3(29), GLI3(31), PIGH(32), CECR5(34), NINJ1(36), DGKG(38), STYXL1(39), DNMT1(43), RASGRP3(44), DEK(45), CLSTN2(46), SCUBE2(50), SLC7A2(52), CSNK1A1(54), SLC16A6(55), VCP(56), MELK(58), TBC1D9(61), KDM4B(67), ASPM(70), ACSM1(76), SKP1(98), ACADVL(78), ADCY1(81), RNF144A(83), BBS4(85),FBXL5(92), UNC119B(95), **TTK(110), AQR(119), MREG(121), VAV3(145), MLPH(164), DNALI1(165), DYRK2(183), YEATS2(200), CCND1(245), PTTG1(252)**
9	MARCH8(1), ASB6(4), GLA(6), CRYZL1(8), IL18R1(24), IRS1(25), CCNE1(27), SOS1(40), CABP2(47), MKL2(51), SMC5(60), ABHD2(65), ORC1(68), JMJD7(86), STK17B(88), PIGH(32), CECR5(34), NINJ1(36), DGKG(38), STYXL1(39), GFRA1(90), **POLDIP3(104), C10orf116(107), BLOC1S1(111), TTC39A(142), PLAGL1(150), TUBGCP4(152), TMSB15B(155), AMFR(163), BLVRA(169), ATPIF1(176), MED13L(192), IGFBP4(198), PJA2(206), MAPT(222), SETD3(229), KIAA0040(243), CENPA(280)**

Appearance number (♯ App.) denotes the number of hypergraph classifiers for which the $d(v)$ of a specific gene is larger than the threshold among the 10 learned models, and thus its maximum value is 10. MI-rank is the rank of the MI value between each gene and the class label. The genes not belonging to top 100 MI rank are bold.

Table 3.5: List of genes with high $d(v)$ in multiple myeloma data, identified by the learned model ($p < 0.05$)

♯ App.	Genes (MI-rank)
10	FSD1(6), HEPACAM(7), CLDN2(13), HSD17B1(53), TDRD3(54), ISOC2(60)
9	LOC100509550(4), ITGAL(17), PREP(19), PGAM2(20), ZMYM1(24), PTDSS2(34), TNNI2(35), QPCT(37), C6orf218(49), SH3KBP1(63), PHLPP1(65), MTMR6(74), FECH(75), RBM45(88), **GGH(102), WHAMM(110),SMAD5OS(125), BPGM(132), NCRNA00208(175), BMP8A(196), GGT7(242), ZACN(258), IFI16(265), CYGB(289), RD3(366), PNKD(375), MOCS3(393), NAT1(581)**

genes with large $d(v)$ but low MI-rank may exert a strong influence on prognostic prediction under the appropriate conditions of other related genes. Moreover, the informative genes repeatedly appear in most of the independently-learned models, indicating that the proposed evolutionary learning method can robustly identify significant hyperedges as prognostic gene modules without the dominant effects of the used prior knowledge. At the same time, the efficiency is enhanced by introducing mutual information to the evolutionary learning of the hypergraph classifiers, without reducing the search space.

Several genes, such as MKI67, CCND1, TTK, PTTG1, CENPA, COX2, and BCL2 have been associated with cancer prognosis in the literature. For example, MKI67 and CCND1 are well-known prognostic markers. They can effectively predict the treatment efficacy of chemotherapy by measuring expression levels of MKI67 and CCND1 (Taneja et al., 2010). TTK and PTTG1 were found to be associated with increased breast cancer risk (Lo et al., 2007). CENPA has also been reported as

Table 3.6: Top 10 gene modules extracted from the learned models in the breast cancer (BC) and multiple myeloma (MM).

BC	
1	[*TTK*[1], *ERBB*2[2], VAX2]
2	[MFAP1, *CCND*1[3], SHCBP1]
3	[GLI3, *PTTG*1[1], SOX11, *TTK*[1]]
4	[C6orf211, NUCB2, *CENPA*[4], ZNF207]
5	[ERLIN2, NEK11, *MKI*67[3], NAT1]
6	[TTC39A, ABCC4, MFAP5, *MKI*67[3]]
7	[*CCND*1[3], HNRNPM, HOXC6, SNTB1, DGKQ]
8	[CTSL2, *MKI*67[3], PDE8B, C16orf42, GLI3]
9	[*MKI*67[3], MARCH8, CABP2, SRSF1, BAG1, RTN2]
10	[PSME4, SOS1, DDX58, ELAVL2, SLC16A6, *CENPA*[4]]
MM	
1	[TEX14, DRAP1, SOX21]
2	[RIOK1, HECW1, CLDN2]
3	[CD58, PAX4, HGFAC, *BCL*2[5]]
4	[ZNF786, *COX*2[6], LOC400128, ANAPC4]
5	[TAX1BP3, *COX*1[6], RPL23A, LOC286149]
6	[SFT2D1, FZD5, TMEM11, YTHDF2, *BCL*2[5]]
7	[EDA, DOC2B, MTMR6, *COX*2[6], GMCL1]
8	[MKNK1, UHRF2, MRPL45P2, TMEM160, ATP5J, *BCL*2[5]]
9	[*COX*1[6], POLE2, SPATA18, C14orf153, NSUN6, SLFN5]
10	[CDK17, TMEM42, *COX*2[6], LZTS2, RAD51, CARS2]

Genes with a superscript number are confirmed to be related to cancer by the following literature: [1] Lo et al., 2007, [2] Schwaetz et al., 1999, [3] Tanega et al., 2010, [4] McGovern et al., 2013, [5] van de Donk et al., 2006, and [6] Ladetto et al., 2005.

Table 3.7: Gene ontology analysis of the clusters from the learned model in breast cancer (p-value < 0.05)

C	Genes	GO ID	Go Terms	p-value
I	MKI67, MARCH8,	GO:0007049	Cell cycle	7.08e-3
	ACADVL, IL18R1,	GO:0006281	DNA repair	9.63e-3
	TTC39A, SPINLW1,	GO:0048589	Developmental growth	1.29e-2
	BTG2, SMC5, HECA,	GO:0006974	Response to	2.25e-2
	GLI3, BAG4, NEK11,		DNA damage stimulus	
	PSMF1, PDE8B, NOLC1,	GO:0008285	Negative regulation of	2.49e-2
	ERLIN2, PQBP1, NAT1		cell proliferation	
II	CENPA, SCUBE2, ANGEL2,	GO:0007338	Single fertilization	1.00e-2
	ASB4, CUTC, DNALI1,	GO:0006281	Cell motion	4.51e-2
	LHX1 C6orf211, ZNF207	GO:0006974	Cellular developmental process	4.62e-2

a significant independent prognostic marker in patients with ER-positive breast cancer (McGovern et al., 2012). In addition, increased COX2 expression is known as an independent adverse prognostic factor in multiple myeloma (Ladetto et al., 2005). BCL2 is also reported to be associated with the response to interferon therapy in multiple myeloma patients (Donk et al., 2006). Thus, high-degree genes identified by evolutionary learning can be prognostic markers for predicting cancer clinical outcomes, since they form hubs in the learned hypergraph structure. Table 3.6 presents an example of hyperedges as potential gene modules influencing on prognosis prediction. In particular we observe that a module involving TTK and PTTG1 appears concurrently from the learned model in the breast cancer. Interest-

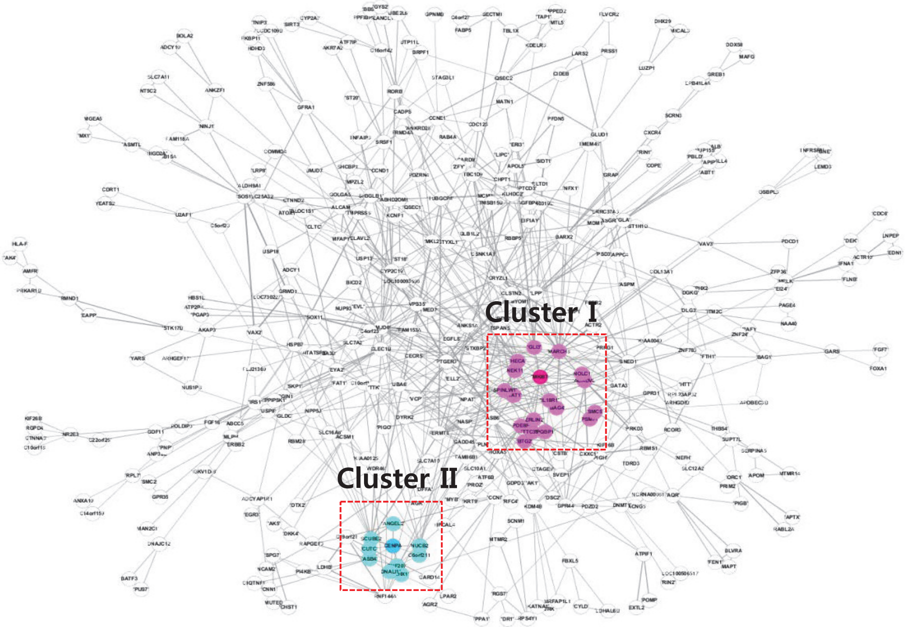


Figure 3.9: Visualization of HCs on breast cancer data. The hypergraph is converted to a normal graph for convenient visualization. This network consists of 422 nodes and 830 edges.

ingly, this finding is consistent with a previous study, in which TTK and PTTG1 act jointly as reproductive risk factors reflecting susceptibility to estrogen exposure for determining breast cancer risk (Lo et al., 2007).

Moreover, the proposed model can be visualized by converting a hyperedge to a clique. Sub-graphs involving genes with large $d(v)$ that are closely related to breast cancer prognosis, such as important prognostic markers, MKI67 and CENPA, are presented in Figure 3.9. In this figure, the cluster is extracted using hypergraph spectral clustering (Zhou et al., 2007), a generalized spectral clustering method (Von, 2007) for hypergraph structures. We also calculated the hypergraph Laplacian L from the learned model, a matrix representing the data variables whose column vectors are eigenvectors of L (Zhou et al., 2007). For clustering, we selected 76 eigenvectors corresponding to eigenvalues below 0.4 from L. Moreover, Table 3.7 shows two gene clusters involved in the network converted from the learned hypergraph with Gene Ontology (GO) analysis (Khatri et al., 2007). The results indicate that genes comprising each cluster have the similar function related to cellular processes. Herein, interpreting the results in this way, we can analyze complex biological phenomena. Thus, the proposed model presents as an alternative method for solving a variety of biomedical problems.

3.5 Summary

We proposed hypergraph classifiers based on evolutionary learning to predict cancer prognoses from complex genetic interactions, using archived data. The learning method evolves a population-based representation of hypergraphs by sequential Bayesian sampling. The Bayesian evolutionary hypergraph model accommodates formal management of model complexity by defining priors on a huge combinatorial search space comprising tens of thousands of genes. Specifically, we controlled the evolutionary search process using two types of prior distributions. One prior guided the compositional variation of the variables in a hyperedge, defined in terms of the mutual information between each genetic variable and the class label. The other was applied on the model size, modulating the degree of a hyperedge and the number of hyperedges in the model.

Cancer prognosis is typically influenced by the combinatorial regulation of multiple genetic factors. By analyzing gene relationships at higher-order levels, we can better predict clinical outcomes in cancer patients. We have demonstrated that higher-order interactions discriminate prognosis more precisely than pair-wise analyzes of single gene relationships. From this viewpoint, we predicted that potential prognostic gene modules could be identified from higher-order gene interactions.

The performance of the proposed method was validated on MAQC-II data. The accuracy of the hypergraph classifiers was similar to that of SVMs and LCSs, and higher than that of naive Bayes classifiers AdaBoost and random forest models. In addition, the MCC of the proposed model was superior to that of existing models. In particular, the MCC score of our model was higher than that of SVMs for multiple myeloma data as 0.34, while the MCC of LCSs was zero for both breast cancer and myeloma datasets. This result indicates that the proposed hypergraph classifiers are robust to imbalanced data, thus more precisely predicting clinical outcomes in cancer patients than existing models. We also compared the performance of the proposed model against two variants of hypergraph classifiers (2-HCs and HCs without using MI as prior). We observe that higher-order relationships are more important for accurately predicting cancer prognosis than pair-wise relationships. Moreover, when hyperedges were generated from information theory, the MCC was improved for both datasets, indicating that searching ability can be enhanced by introducing problem-specific knowledge to the prior in the evolutionary learning process. Furthermore, the interpretable structures of hypergraph classifiers proved useful for analyzing complex biological phenomena. That is, the proposed model presents as an alternative method for solving a variety of biomedical problems. Such contributions will greatly assist toward developing a personalized and refined therapy.

Chapter 4

Hypergraph-based Models for Constructing Higher-Order miRNA-mRNA Interaction Networks in Cancer

4.1 Overview

Dysregulation of genetic factors such as microRNAs (miRNAs) and mRNAs has been widely shown to be associated with cancer progression and development. In particular, miRNAs and mRNAs cooperate to affect biological processes, including tumorigenesis. The complexity of miRNA-mRNA interactions presents a major barrier to identifying their co-regulatory roles and functional effects. Thus, by computationally modeling these complex relationships, it may be possible to infer the gene interaction networks underlying complicated biological processes.

In this chapter, we introduce a data-driven model for identifying cancer stage-

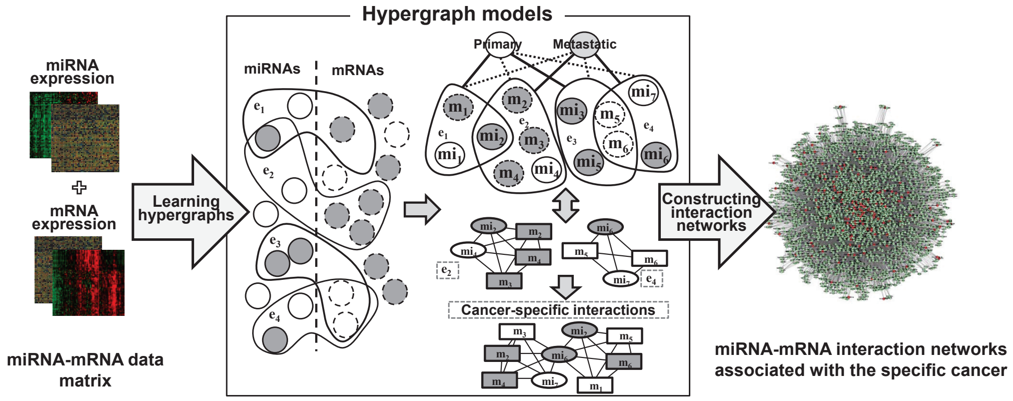


Figure 4.1: Overview of the hypergraph-based models for constructing higher-order miRNA-mRNA interaction networks at a specific cancer stage. Solid and dotted circles denote miRNAs and mRNAs, respectively. Closed curves denote hyperedges (i.e. modules). In the conventional graph representation (two graphs in the right-bottom of the central box of the figure), ellipses and boxes denote miRNAs and mRNAs, respectively. Grey and white indicate respective high and low gene expression levels.

specific interactions that reflects the high-order relationships between miRNAs and mRNAs (Figure 4.1) (Kim et al., 2012b, 2013b). The proposed model is a hypergraph comprising numerous hyperedges, representing the multi-variable combinations corresponding to miRNAs and mRNAs. Each hyperedge is formally defined as cancer-stage specific statistical figures, and thus our model can deal with real-valued data without discretization. The weight of a hyperedge reflects the strength of the higher-order dependency among the variables of the hyperedge. Therefore, each hyperedge potentially behaves as a gene module. The model explicitly constructs a complex interaction network from many such gene modules. The model is learned by finding a highly-discriminate hypergraph structure from expression profiles using data relevant to a certain stage of prostate cancer.

The learning process involves the iteration of two learning phases; structure and parameter. The structure learning phase constructs a hypergraph of putative hyperedges for discovering potential gene interactions, from a huge feature space represented by the combinations of many miRNAs and mRNAs. Because the miRNA-mRNA interactions are intractably complex, we adopt an evolutionary strategy based on an information theoretic co-regulatory measure, called mutual information. This strategy is used to select genetic variables for generating hyperedges. During the parameter learning phase, the hypergraph is refined by updating the weights of the hyperedges (representing higher-order miRNA-mRNA modules). To this end, we employ a gradient descent method similar to the back-propagation algorithm for learning artificial neural networks. The learned model is then converted into a network structure reflecting the cooperative higher-order gene activities by connecting the extracted hyperedges. Data-driven learning allows the model to build new miRNA-mRNA interaction networks which display the hidden properties of primary and metastatic prostate cancers from a given dataset, which are not known *a priori*.

We construct cancer stage-specific miRNA-mRNA interaction networks reflecting their higher-order relationships using the MSKCC Prostate Oncogenome Project dataset from the model (Taylor et al., 2010). We demonstrate that the proposed model can build several biologically significant miRNA-mRNA interaction networks, including potential modules associated with primary and metastatic prostate cancer. Moreover, cancer-related miRNAs and genes dominate the identified interactions. Some of these interactions, such as hsa-miR-1, hsa-miR-133a, hsa-miR-143, hsa-miR-145, hsa-miR-221, hsa-miR-222, act as hubs in the constructed networks.

We also confirm the biological relevance of the constructed networks through literature review and functional analysis.

4.2 Analyzing Relationships between miRNAs and mRNAs from Heterogeneous Data

Recently, miRNAs have caused great excitement as diagnostic and therapeutic signatures of prostate cancer (Coppola et al., 2010; Gordanpour et al., 2012; Watahiki et al., 2011; Schaefer et al., 2010). They play important roles in cancer pathogenesis, including disease onset, progression, and metastasis, by regulating the stability and translation efficiency of their target mRNAs. Thus, the functional relationships between miRNAs and mRNAs should be elucidated to identify key transcriptional circuits involved in cancer regulation. However, analyzing higher-order miRNA-mRNA relationships is rendered as a challenging problem due to the complexity of their interactions.

Several studies have attempted to identify groups of coherent miRNAs and mRNAs that cooperate in biological processes from heterogeneous data sources via various computational approaches, including probabilistic methods (Yoon and Micheli, 2005; Huang et al., 2006; Joung et al., 2007; Joung and Fei, 2009; Liu et al., 2009a; Bonnet et al., 2010a,b; Liu et al., 2010), rule-based learning (Tran et al., 2008; Liu et al., 2009b), matrix factorization (Zhang et al., 2011), and statistical methods (Peng et al., 2009; Nunez-Iglesias et al., 2010; Lu et al., 2010; Zhang et al., 2012c). These approaches have simplified complex biological mechanisms by systematically analyzing the relationships between genetic elements at the genome level.

Typically, however, bi-relationships between only two factors are assumed in many previous studies (Yoon and Micheli, 2005; Liu et al., 2009b; Zhang et al., 2011;

Peng et al., 2009; Nunez-Iglesias et al., 2010; Lu et al., 2010; Zhang et al., 2012c). Such restrictions are unsuitable for complex genetic interactions because information is lost under the assumption, and biological regulation is controlled by the interaction of multiple genetic components. Many studies have also investigated miRNA-mRNA regulatory interactions using biological information, especially miRNA-target information (Yoon and Micheli, 2005; Huang et al., 2006; Joung et al., 2007; Joung and Fei, 2009; Liu et al., 2009a; Tran et al., 2008; Liu et al., 2009b; Zhang et al., 2011; Peng et al., 2009; Nunez-Iglesias et al., 2010). Biological information reduces the number of false positives, since it provides the predictive model with prior knowledge. In contrast, unknown or hidden interactions not involved in the prior knowledge may be difficult to identify from this information.

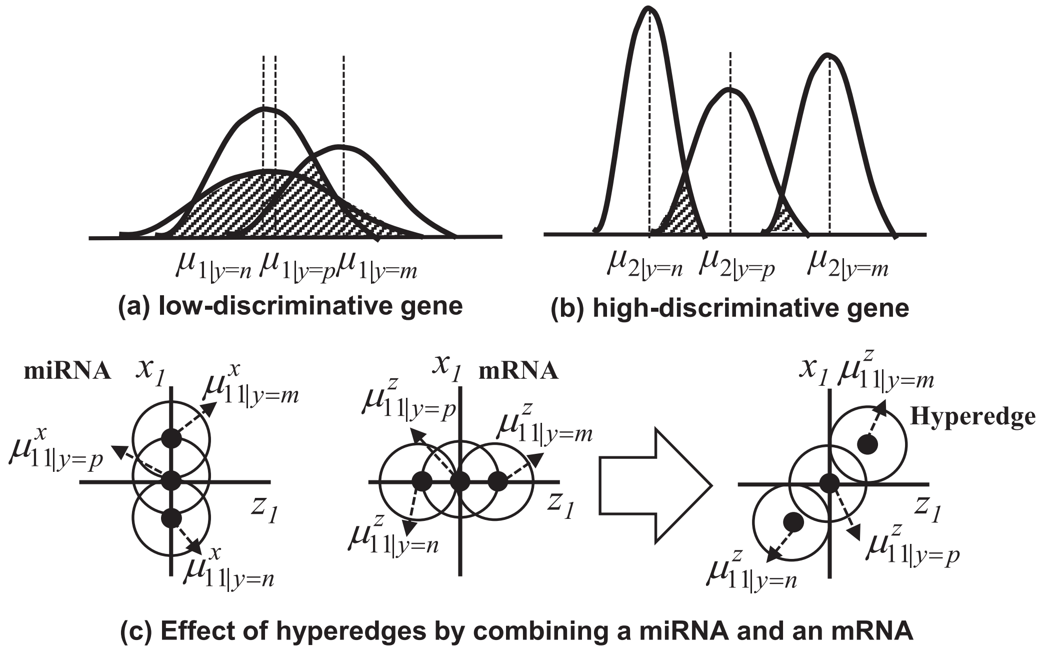
To avoid this problem, some probabilistic models which infer miRNA-mRNA modules from expression profiles only, without relying on target information, have been proposed (Bonnet et al., 2010a,b; Liu et al., 2010). Bonnet's model, called LeMoNe (Bonnet et al., 2010a,b) consists of two major steps; the generation of gene clusters based on a feature-sample co-clustering method, and the inference of regulatory modules from generated clusters and regulators based on probabilistically optimized trees. In the clustering approach of Bonnet's method, gene regulatory modules underlying a specific cancer stage are not easily identified. Liu's approach infers functional miRNA regulatory modules using Correspondence Latent Dirichlet Allocation (Corr-LDA) (Liu et al., 2010). The Corr-LDA based model requires discretized data. Since the Corr-LDA model infers probability distributions from latent variables, moreover, miRNAs can be annotated to any functional modules, while mRNAs are restricted to the miRNA-inferred modules.

4.3 Hypergraph-based Models for Identifying miRNA-mRNA Interactions

4.3.1 Hypergraph-based Models

A hypergraph-based model characterizes complex interactions among many genetic factors using hypergraph structures. A hypergraph generalizes the edge concept to a hyperedge by which more than two variables can be connected simultaneously (Zhang, 2008; Kim et al., 2010). As such, it is suitable for representing higher-order relationships among heterogeneous features (e.g. miRNAs and mRNAs). In our model, a hyperedge contains two or more variables corresponding to miRNAs and mRNAs, weighted by the strength of the higher-order dependency among its elements for each class (where the class denotes a specific cancer stage). Thus, each hyperedge implies a set of miRNA-mRNA modules associated with a certain stage of cancer. The proposed model therefore facilitates the construction of higher-order miRNA-mRNA interaction networks among a population of candidate gene modules related to a specific cancer stage.

A hypergraph-based model H is formally defined as a triple $H = (X, Z, E)$ where X, Z, and E denote the sets of miRNAs, mRNAs, and hyperedges, respectively. A hyperedge is represented by a set of statistical values, including mean and covariance for the class label corresponding to a cancer stage. The mean gene expression values differ widely among the class labels, implying that gene expression depends on cancer progression, as shown in Figure 4.2. The hyperedge approach enhances the discriminative capability by combining miRNAs and mRNAs (Figure 4.2). Given an expression dataset with N instances $D = \{d^{(n)}\}_{n=1}^{N} = \{\mathbf{x}^{(n)}, \mathbf{z}^{(n)}, y^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}^{(n)}$ and $\mathbf{z}^{(n)}$ are real-valued vectors of miRNA and mRNA expressions in the n-th instance, and y is an element of a cancer stage set Y, the i-th hyperedge e_i



(a) low-discriminative gene

(b) high-discriminative gene

(c) Effect of hyperedges by combining a miRNA and an mRNA

Figure 4.2: Biological meaning of mean and variance used in representing a hyperedge. Panels (a) and (b) illustrate how the means and variances differ between low and high discriminative genetic factors. A gene is low-discriminative when the means are similar at each disease stage but the variances are large (where n, p, and m denote normal, primary, and metastatic stage, respectively). Panel (c) illustrates the enhanced discriminative capability of a hyperedge involving two genetic factors. By comparing the discriminative capability of each miRNA or mRNA, the discrimination capability of the hyperedge is enhanced.

contains the mean vectors and the covariance of its miRNAs and mRNAs for the given cancer stage:

$$e_i = \left\{ \begin{array}{c} e_{i|y=y_1} \\ \dots \\ e_{i|y=y_{|Y|}} \end{array} \right\} = \left\{ \begin{array}{c} (\mu_i, \Sigma_i)_{|y=y_1} \\ \dots \\ (\mu_i, \Sigma_i)_{|y=y_{|Y|}} \end{array} \right\}, \tag{4.1}$$

$$\mu_i = (\mu^x_{i1}, ..., \mu^x_{il}, \mu^z_{i1}, ..., \mu^z_{im}) \text{ and } l + m = |e_i| \tag{4.2}$$

where μ^x_{ij} and μ^z_{ik} denote the means calculated from the expression profiles of the j-th miRNA and the k-th mRNA, respectively, in the i-th hyperedge (whose elements comprise l miRNA and m mRNAs). l and m are called the degrees of miRNA and mRNA of the hyperedge, respectively. By the definition of a hyperedge, each hyperedge has $|Y|$ mean vector/covariance pairs, and $|Y|$ weights. The hypergraph-based model is considered as a population of hyperedges. Given a gene expression profile ($\mathbf{x}$, $\mathbf{z}$), the cancer stage of the profile is classified as y^*, for which the summation of the expected values (the products of the hyperedge weight and the probability of ($\mathbf{x}$, $\mathbf{z}$) matching the hyperedge), is highest among the elements of Y. "($\mathbf{x}, \mathbf{z}$) matches $e_i|y$" means that ($\mathbf{x}$, $\mathbf{z}$) has similar expression values to ones of the i-th hyperedge with respect to the genetic variables involved in $e_i|y$ at cancer stage y, and we introduce a Gaussian kernel into the hyperedge to calculate the matching probability of ($\mathbf{x}$, $\mathbf{z}$) and $e_{i|y}$, $P(u = 1|\mathbf{x}, \mathbf{z}, e_{i|y})$. The matching probability is calculated by the normalized subdimensional distance between $e_{i|y}$ and ($\mathbf{x}$, $\mathbf{z}$):

$$P(u = 1|x, z, e_{i|y}) = \exp\left\{-\beta d(x, z, e_{i|y})\right\}, \tag{4.3}$$

$$d(x, z, e_{i|y}) = \frac{1}{|e_i|} \left\{ \sum_{j=1}^{l} \frac{\left(x_{ij} - \mu_{ij}^{x}\right)^2}{\left(\sigma_{ij|y}^{x}\right)^2} + \sum_{k=1}^{m} \frac{\left(z_{ik} - \mu_{ik}^{z}\right)^2}{\left(\sigma_{ik|y}^{z}\right)^2} \right\}^{\frac{1}{2}}, \tag{4.4}$$

where u=1 denotes that (**x**, **z**) matches $e_{i|y}$, $\sigma_{ij|y}^{x}$ and $\sigma_{ij|y}^{z}$ are the standard deviations of x_{ij} and z_{ik} (the j-th miRNA and k-th mRNA, respectively) in the i-th hyperedge for a given y, and β is a constant for adjusting the probability. Larger β implies smaller matching probability, and therefore a smaller number of hyperedges influence on classifying the data. Specifically, the cancer stage y^* of (**x**, **z**) is computed as follows:

1. Calculate $c_{y'}$, the sum of the expected values for each y' in Y over all hyperedges of H:

$$c_{y'} = \sum_{i=1}^{|H|} w(e_{i|y=y'}) P(u = 1 | x, z, e_{i|y=y'}) \tag{4.5}$$

   where $|H|$ denotes the number of hyperedges and $w(e_{i|y})$ is the weight of $e_{i|y}$, explained in the next subsection.

2. Predict the cancer stage as y^*:

$$y^* = \arg\max_{y' \in Y} c_{y'} \,. \tag{4.6}$$

In terms of distance-based connectionist models, our model is related to radial basis function networks (RBFNs) (Buhmann, 2003). Whereas RBFNs use kernelized distance for all variables, the proposed hypergraph model uses the probability derived from the subdimensional distance on the projected space corresponding to each hyperedge. Unlike RBFNs, therefore, the hypergraph model can detect embedded subpatterns reflecting higher-order relationships among the components. Because these embedded subpatterns influence the classification, we can intuitively analyze the complex interactions of genetic factors that contribute to classifying a specific cancer stage.

4.3.2 Learning Hypergraph-based Models

The proposed model learns by finding a hypergraph structure with high discriminative capability at a specific cancer stage. This is achieved by maximizing the conditional likelihood for a model H and the gene expression profiles and a log function is adopted for convenience. To minimize the error of classifying the cancer stage, $E_{D,H}$, the log conditional likelihood is maximized by least mean square criteria using (4.5) and a sigmoidal function:

$$\begin{aligned}
&H* = \arg\max_H \log \prod_{n=1}^{N} p(y^{(n)}|x^{(n)}, z^{(n)}, H) = \arg\max_H \sum_{n=1}^{N} \log p(y^{(n)}|x^{(n)}, z^{(n)}, H) \\
&\equiv \arg\max_H \sum_{n=1}^{N} \delta(y^{(n)}, y'_H) = \arg\min_H E_{D,H} \qquad (4.7)\\
&\approx \arg\min_H \sum_{n=1}^{N} \sum_{y' \in Y} \left(\delta(y^{(n)}, y') - P(y'|x^{(n)}, z^{(n)}, H)\right)^2
\end{aligned}$$

s.t.

$$P(y'|x, z, H) = \left\{1 + \exp\left(c_{y'} - \frac{1}{|Y|} \cdot \sum_{y \in Y} c_y\right)\right\}^{-1}$$

where ($\mathbf{x}^{(n)}$, $\mathbf{z}^{(n)}$) denotes the n-th miRNA-mRNA expression and $y^{(n)}$ is the cancer stage of the example. y'_H is the label predicted by H and $\delta(y^{(n)}, y'_H)$ is an indicator function, equal to 1 if $y^{(n)}$ equals y'_H, and 0 otherwise. To enhance the classification accuracy, it is essential that the population comprises hyperedges with high discriminative capability, and the hyperedge weights must be refined to minimize (4.7) in the generated hypergraph.

To meet these requirements, the learning iterates two phases: structure learning and parameter learning. The structure learning constructs a hypergraph from hyperedges that identify potential miRNA-mRNA modules. The weights of the hyperedges are updated to minimize the classification error of the generated gene module population during the parameter learning phase. Because the hypergraph model represents a huge combinatorial feature space (size $2^{|x|+|z|}$) of many miRNAs and mRNAs, exhaustively searching for the optimal population is infeasible. Instead we adopt an evolutionary learning method based on information-theoretic criteria to generate putative hyperedges for the structure learning.

We assume that a hyperedge consisting of strongly interactive miRNAs and mRNAs is highly discriminative for classification in this study. Mutual information is used as a co-regulatory measuring criterion for efficiently selecting genes for hyperedge generation. Mutual information (MI) is an information-theoretic measure that specifies the degree of conditional independency between two random variables. When a genetic factor more strongly determines the cancer stage, the MI between the gene and the cancer stage is increased. A hyperedge is generated by probabilistically selecting miRNAs and mRNAs, and the MI between each gene and the class label determines the probability of selecting the genes. The probability $P_I(X_i)$ of selecting the i-th gene X_i is defined such that miRNAs or mRNAs with high MI are selected more frequently:

$$P_I(X_i) = \frac{\{I(X_i;Y)\}^\eta}{\sum\limits_{X_i \in X} \{I(X_i;Y)\}^\eta}, \tag{4.8}$$

where denotes the MI between the i-th genetic factor and the cancer stage, and η is a nonnegative constant that regularizes the influence of MIs on the gene selection. When η is zero, all variables may be selected with equal probability. Once the hyperedges have been generated, the mean vectors and covariance of the hyperedges are calculated from the training dataset. To identify putative strongly-interacting miRNA-mRNA modules, the initial weight of the *i*-th hyperedge is computed using the variances of each genetic factor and the multivariate MI (Kraskov et al., 2004) among all variables, including the class label involved in the hyperedge. A gene with a particular mean expression value but small variance likely possesses higher discriminative capability than one with larger variance. Moreover, by the definition of MI, large multivariate MI implies more relationships among the genes. Thus the initial weight of a hyperedge is defined as

$$w_0(e_{i|y}) = \kappa \cdot I(e_i) + \sum_{x_{ij} \in e_i} \frac{1}{\sigma^2_{ij_{|y}}} \tag{4.9}$$

s.t.

$$\begin{aligned} I(e_i) &= I(X_{i1,};..;X_{ik};Y) = I(X_{i1,};..;X_{ik}) - I(X_{i1,};..;X_{ik}|Y) \\ &= I(X_{i1,};..;X_{ik}) - E_Y(I(X_{i1,};..;X_{ik})|Y)\ , \end{aligned}$$

where k is the number of variables of e_i and κ denotes the ratio of the variance to MI. In the parameter learning phase, the weights of the hyperedges are updated using the gradient descent method for all training data. The aim is to minimize the error in terms of the classification probability in (4.3) and the matching probability in (4.7):

$$w_t(e_{i|y}) = \Delta w_{t,i|y} + w_{t-1}(e_{i|y}), \tag{4.10}$$

$$\Delta w_{t,i|y} = \frac{\gamma}{t} P(y|\mathbf{x}, \mathbf{z}, H)\,(1 - P(y|\mathbf{x}, \mathbf{z}, H))\,(\delta(\tilde{y}, y) - P(y|\mathbf{x}, \mathbf{z}, H)) \cdot P(u = 1|\mathbf{x}, \mathbf{z}, e_{i|y}),$$

where $\tilde{y}$ is the real cancer stage of a miRNA-mRNA expression sample, and t and γ denote the epoch number in the parameter learning and the parameter learning rate, respectively. The epoch is the number of weight updates for the built hypergraph during parameter learning, and γ controls the extent of weight change during parameter learning. Thus, the weight becomes high when the hyperedge consists of miRNAs and mRNAs with strong higher-order interactions and when the variances of the gene variables are small at all cancer stages. Following parameter learning, low weighted hyperedges are removed from the population, and the next structure learning step is performed. To prevent the removal of highly discriminating hyperedges, the number of replaced hyperedges decreases to a specific value as the iterations proceed, as follows:

$$R_t = \frac{R_{max} - R_{min}}{\exp(t)} + R_{min}, \tag{4.11}$$

where t is the iteration number of the structure learning phase, and R_{max} and R_{min} denote the maximum and minimum number of replaced hyperedges, respectively. Therefore, the number of replaced hyperedges consecutively decreases as the structure learning proceeds, while high-discriminative modules are preserved. The algorithm for learning the hypergraph-based model is presented in Figure 4.3.

4.3.3 Building Interaction Networks from Hypergraphs

We construct a higher-order miRNA-mRNA interaction network at a specific cancer stage from the learned model. When analyzing complex biological networks based on graph mining, frequently occurring subgraphs in the networks are generally regarded as important building blocks which are merged to create the functional network (Hu et al., 2005; Mason and Verwoerd, 2007; Yan et al., 2007; Ramadan et al., 2010). Since a high-weight hyperedge corresponds to a significant subgraph

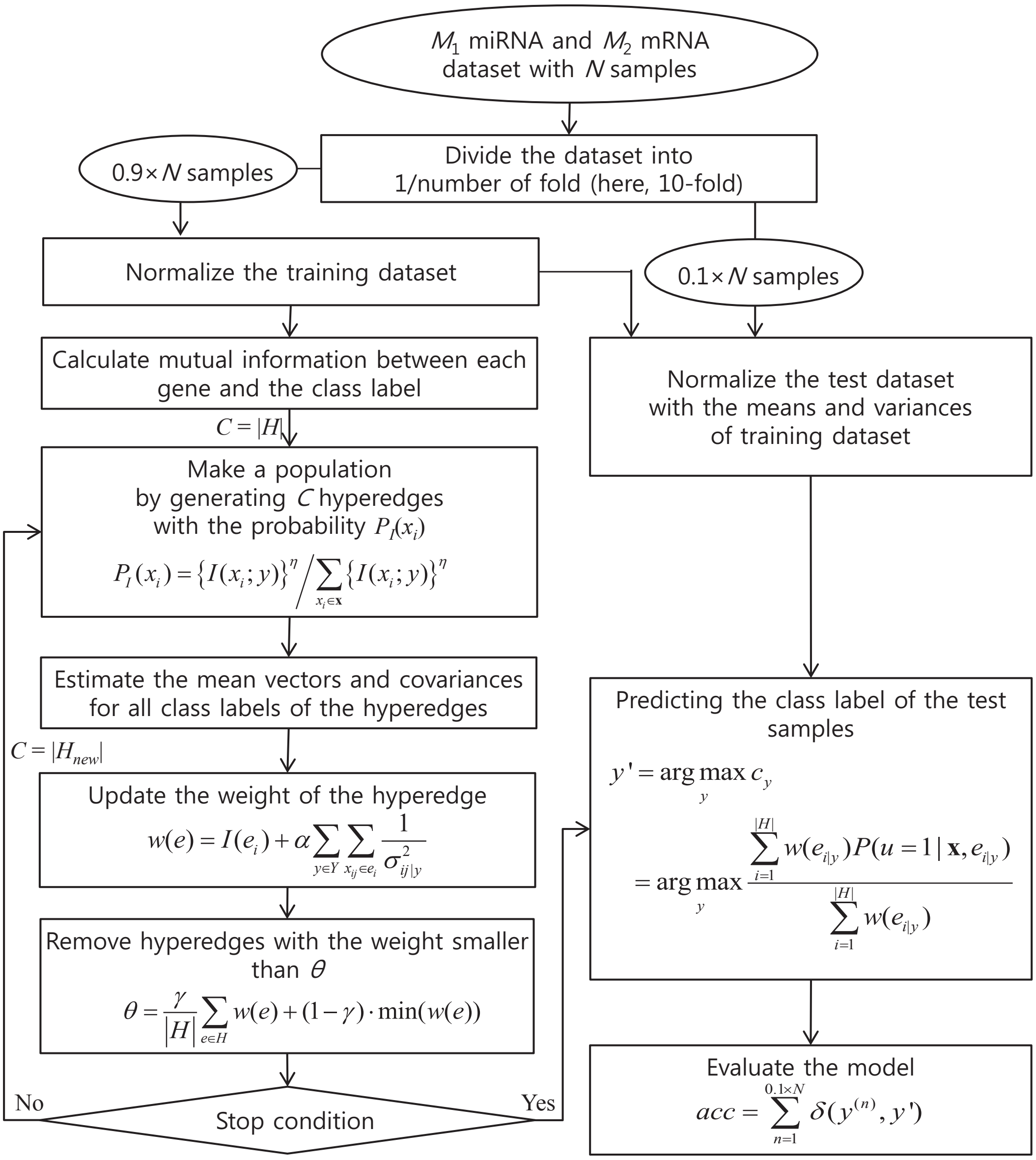

M_1 miRNA and M_2 mRNA dataset with N samples
Divide the dataset into 1/number of fold (here, 10-fold)
0.9× N samples
0.1× N samples
Normalize the training dataset
Calculate mutual information between each gene and the class label
$C = |H|$
Make a population by generating C hyperedges with the probability $P_I(x_i)$
$P_I(x_i) = \{I(x_i;y)\}^\eta \Big/ \sum_{x_i \in \mathbf{x}} \{I(x_i;y)\}^\eta$
Estimate the mean vectors and covariances for all class labels of the hyperedges
$C = |H_{new}|$
Update the weight of the hyperedge
$w(e) = I(e_i) + \alpha \sum_{y \in Y} \sum_{x_{ij} \in e_i} \frac{1}{\sigma^2_{ij|y}}$
Remove hyperedges with the weight smaller than θ
$\theta = \frac{\gamma}{|H|} \sum_{e \in H} w(e) + (1-\gamma) \cdot \min(w(e))$
No
Stop condition
Yes
Normalize the test dataset with the means and variances of training dataset
Predicting the class label of the test samples
$y' = \arg\max_y c_y$
$= \arg\max_y \frac{\sum_{i=1}^{|H|} w(e_{i|y}) P(u=1 \mid \mathbf{x}, e_{i|y})}{\sum_{i=1}^{|H|} w(e_{i|y})}$
Evaluate the model
$acc = \sum_{n=1}^{0.1 \times N} \delta(y^{(n)}, y')$

Figure 4.3: Algorithm for learning the hypergraph-based model

reflecting a higher-order relationship among genetic variables, the interaction network is constructed by connecting cliques sharing common genes. A hyperedge is assigned separate weights for each cancer stage and it is merged into the graph of the highest weighted cancer stage. Formally, a cancer-stage and a cancer stage-specific interaction network $G_{|y'}$ =(V, E), where V and E denote a vertex set and an edge set, respectively, is constructed by merging the hyperedges as follows (where y' is the class label with the largest weight value):

$$G_{|y'} = G_{|y'} \cup C_i, \tag{4.12}$$

$$y' = \arg\max_{y \in Y} \left\{ w(e_{i|y}) \right\}, \tag{4.13}$$

and C_i is a clique corresponding to the i-th hyperedge e_i (Figure 4.4). This dividing and remerging approach enables the constructed interaction networks to be easy-to-visualized without impairing the higher-order property of the model since the weight of edges in the constructed networks are derived from the hyperedge weights reflecting the strength of the higher-order interaction.

4.4 Constructing miRNA-mRNA Interaction Networks Based on Higher-Order Relationships

4.4.1 Data and Experimental Settings

The clinical heterogeneity of prostate cancer, coupled with its high prevalence, raises challenges in the management of newly diagnosed patients as well as those with metastatic disease. Specifically, prostate cancer shows enormous biological

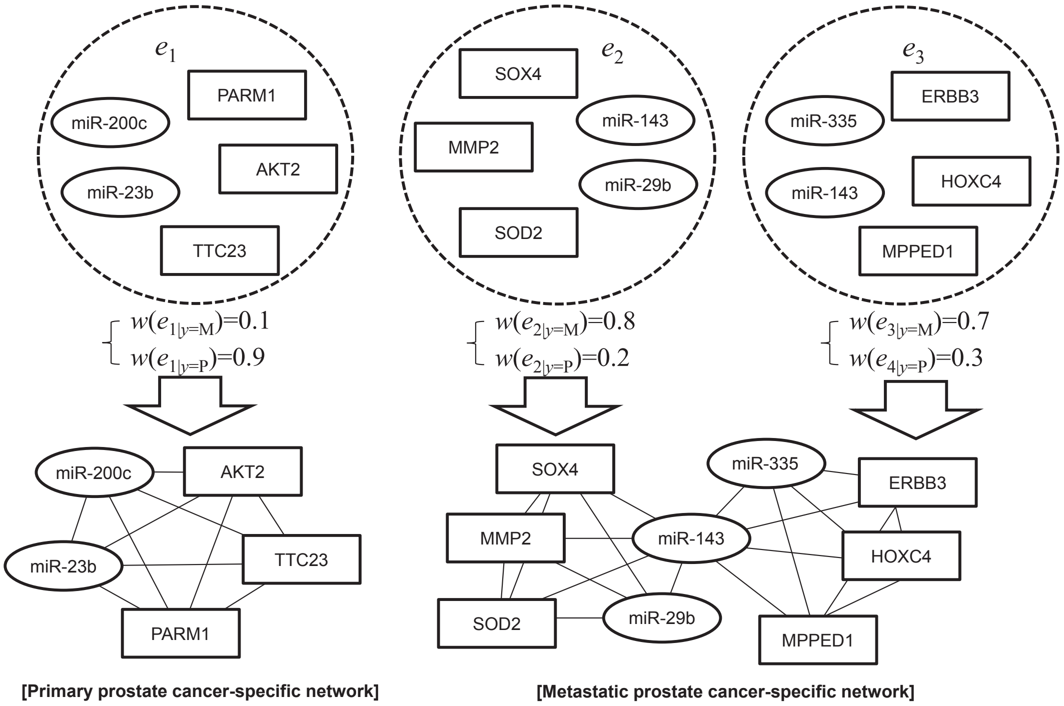

e_1
PARM1
miR-200c
AKT2
miR-23b
TTC23
e_2
SOX4
miR-143
MMP2
miR-29b
SOD2
e_3
ERBB3
miR-335
HOXC4
miR-143
MPPED1
$w(e_{1|y=\mathrm{M}})=0.1$
$w(e_{1|y=\mathrm{P}})=0.9$
$w(e_{2|y=\mathrm{M}})=0.8$
$w(e_{2|y=\mathrm{P}})=0.2$
$w(e_{3|y=\mathrm{M}})=0.7$
$w(e_{4|y=\mathrm{P}})=0.3$
miR-200c
AKT2
miR-23b
TTC23
PARM1
SOX4
MMP2
miR-143
SOD2
miR-29b
miR-335
ERBB3
HOXC4
MPPED1
[Primary prostate cancer-specific network]
[Metastatic prostate cancer-specific network]

Figure 4.4: Procedure of converting a hypergraph to cancer stage-specific interaction networks. 'P' and 'M' denote metastatic and primary prostate cancer, respectively.

heterogeneity, with some patients dying of metastatic disease within 2-3 years of diagnosis whereas others can live for 10-20 years with organ-confined disease, likely a reflection of underlying genomic diversity. Herein, understanding prostate cancer mechanisms requires integrated large-scale cancer genomic projects which can provide new insights into the molecular classification of cancers. In particular, miRNAs have been recognized as the key regulator of gene expression in prostate cancer. Thus, the integrated analysis of miRNA and mRNA expression on a genome-wide level can offer more informed clinical decision-making and novel therapeutic

Table 4.1: Parameter settings for experiments

Parameters	Values	Parameters	Values
♯ of miRNAs	3	♯ of mRNAs	5
♯ of modules	variable	β in (4.3)	1.0
Epochs of structure learning	100	Epochs of parameter learning	20
η in (10)	1.0	κ in (4.9)	1.0
γ in (13)	1.0	R_{max}, R_{min}	0.9, 0.5

targets.

In this study, miRNA and mRNA expression profiles obtained from the MSKCC Prostate Oncogenome Project (Taylor et al., 2010) were matched at three stages of prostate cancer. The dataset contains 373 miRNAs and 19,780 mRNAs from 27 normal, 98 primary and 13 metastatic stages. During preprocessing, sample-wise and feature-wise normalization was conducted, and miRNAs and mRNAs were separately normalized. The experimental parameter settings are listed in Table 4.1. The parameters are those yielding optimal performance in empirical experiments. A hypergraph can include hyperedges with different number of genetic variables but we fixed the number of variables for all hyperedges of a hypergraph in this study.

4.4.2 Classification Performance

Classification performance was evaluated using three standard classification models; support vector machines (SVMs) with the 2nd polynomial kernel and sequential minimal optimization (SMO), *k*-th nearest neighbor classifiers (*k*-NNs), and naive

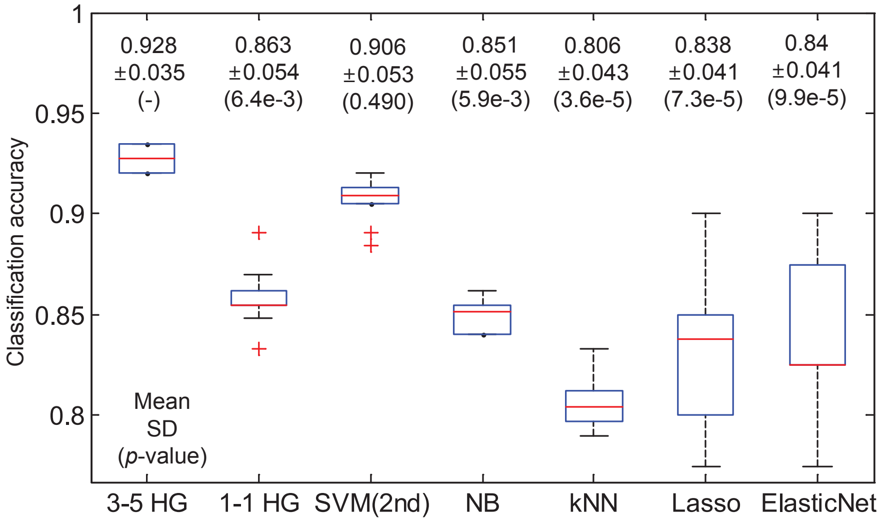


Figure 4.5: Boxplots of classification accuracy on the test set. it m-n HG denotes the hypergraph-based model whose all hyperedges embody m miRNAs and n mRNAs. All results are averaged after 10 runs by 10-fold cross validation. P-values are calculated using t-test of our model and other models.

Bayes classifiers (NBs) implemented in Weka (Hall et al., 2009). The MATLAB algorithms lasso and elastic net (α=0.5) were also used. All results were averaged over 10 experiments. Figure 4.5 presents the classification accuracy of our model compared to other models. As revealed by the p-values of the t-test, the proposed hypergraph-based model competes on-par with SVMs and outperforms the k-NN, NB and Lasso-based methods. In addition, by comparing the results of 3-5 HG (a hypergraph model whose hyperedges consist of three miRNAs and five mRNAs) and 1-1 HG, we observe that higher-order relationships are more important for discriminating cancer stages than pair-wise relationships between a single miRNA and mRNA.

4.4.3 Model Evaluation

The proposed hypergraph-based learning method is evaluated on simulation data for verifying whether the method finds true solutions. The data consist of 500 instances with 7 variables whose mean is zero and the class label of each instance is determined as follows:

$$x_i \sim N(0,1),\ \ 1 \leq i \leq 7$$

$$c^{(n)} = \begin{cases} 1, & if\, x_2 > 2\ \wedge\ x_3 > 2\ \wedge\ x_4 > 2 \\ 2, & if\, x_5 < -2\ \wedge\ x_6 < -2\ \wedge\ x_7 < -2 \\ \quad 3, & otherwise \end{cases}, \tag{4.14}$$

where x_i and $c^{(n)}$ denote the i-th random variable and the class label of the n-th instance. Table 4.2 illustrates the classification accuracy and predefined modules in the learned model. The accuracy is averaged after 10 experiments by 10-fold cross

Table 4.2: Verification result on the simulation dataset

Models	SVM	DT	kNN	HG	Module 1	Module 2
Accuracy	0.956	0.886	0.93	0.956	10	10
±SD	±0.002	±0.004	±0.006	±0.003	-	-

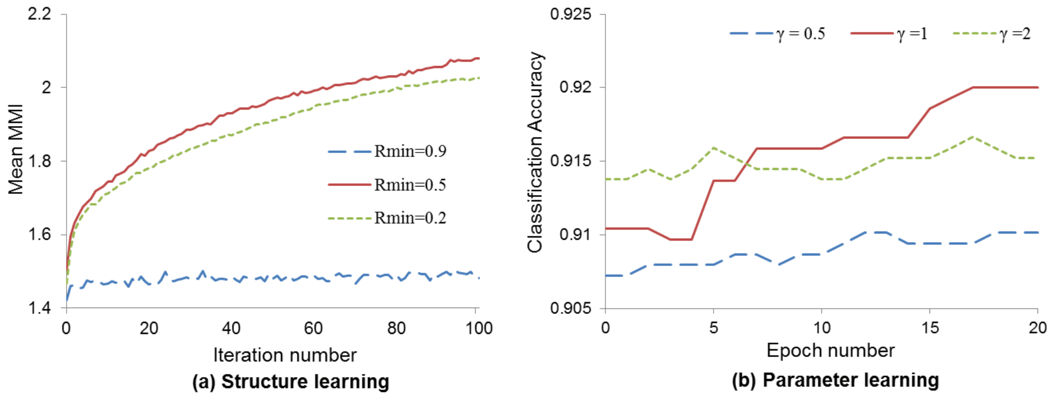


Figure 4.6: Learning curves in the structure and the parameter learning phases. As the performance measure, we used mean multivariate mutual information (MMI) of all hyperedges in the model for the structure learning and accuracy on 10-fold cross validation for the parameter learning. *Rmax* is fixed as 0.9 in (a) and γ is a learning rate for the parameter learning in (b). All results are averaged on 10 experiments of 10- fold cross validation.

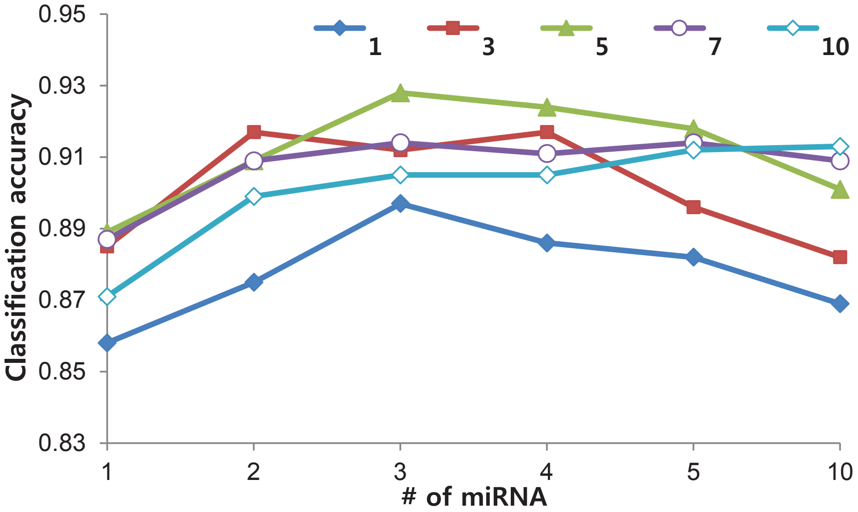
validation, and each hypergraph includes 20 hyperedges with four variables. In Table 2, Module 1 and 2 means the number of case when there exist hyperedges involving a predefined-set 1 (x_2, x_3, x_4) and 2 (x_5, x_6, x_7) in a learned hypergraph. Because we conducted 10-fold cross validation, the maximum values of Module 1 and 2 are ten. Therefore, we indicate that our method can find true solutions from small combinatorial spaces, considering the accuracy and the number of found variable modules.

Figure 4.6 presents two learning curves under various conditions of the structure (a) and the parameter (b) learning phases. As the measure for structure learning, we used mean multivariate mutual information (MMI) of all hyperedges in the model because the goal of the structure learning is to find the significant higher-order cancer-specific gene interaction modules, and an MMI is the measure reflecting the strength of interactions among genetic factors in the hyperedges considering the stage of cancer. On the other hand, classification accuracy is used as the measure for the parameter learning phase since the weight for each cancer stage is updated to minimize the error in the phase. Figure 4.6 (a) presents the increase of mean MMI under various *Rmin* which is the minimum ratio of the hyperedges replaced in the iteration, and plays a role of the structure learning rate. We indicate that too large an *Rmin* causes low MMI by replacing too many hyperedges and too small an *Rmin* leads slow increase of the MMI from Figure 4.6 (a). Figure 4.6 (b) presents similar results to (a) with respect to the effect of learning rate γ.

Moreover, Figure 4.7 shows the classification accuracy according to the number of genetic factors in the hyperedges. The classification accuracy is the best when a hypergraph consists of hyperedges with three miRNAs and five mRNAs. We indicate that small number of genetic variables show worse performance because various processes of prostate cancer is influenced on the complex interactions among many features. Furthermore, the accuracy of the hypergraphs including hyperedges with more than ten genetic variables is low since the models consist of too specific information and thus have the low generalization property.

Figure 4.8 shows that the proposed learning method can stably extract significant genetic factors despite its random selection approach. We define a measure as the number of appearance of a gene in the model, $A(x_i)$, for verifying the stability of the model as follows:

$$A(x_i) = \sum_{m=1}^{100} \delta(x_i, H_m)$$



mRNA \ miRNA	1	2	3	4	5	10
1	0.858(0.015)	0.875(0.009)	0.897(0.012)	0.886(0.019)	0.882(0.01)	0.869(0.01)
3	0.885(0.015)	0.917(0.011)	0.912(0.01)	0.917(0.005)	0.896(0.01)	0.882(0.011)
5	0.889(0.011)	0.909(0.011)	**0.928(0.007)**	0.924(0.006)	0.918(0.009)	0.901(0.008)
7	0.887(0.01)	0.909(0.005)	0.914(0.006)	0.911(0.005)	0.914(0.008)	0.909(0.008)
10	0.871(0.007)	0.899(0.007)	0.905(0.011)	0.905(0.007)	0.912(0.008)	0.913(0.011)

accuracy (±SD)

Figure 4.7: Classification accuracy according to the number of miRNA and mRNA in the hyperedges. The classification accuracy is the best when a hypergraph consists of hyperedges with three miRNAs and five mRNAs. All results are averaged on 10 experiments of 10-fold cross validation.

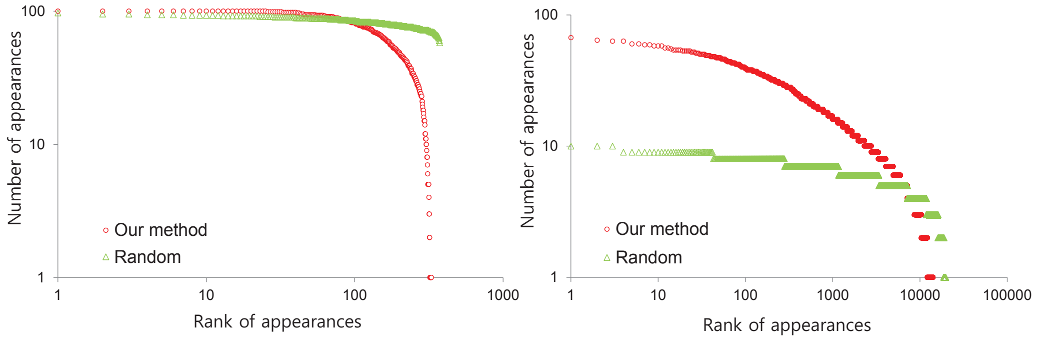
$$\delta(x_i, H_m) = \begin{cases} 0 \mathrm{if} x_i \mathrm{isnotinvolvedin} H_m \\ 1 \ \mathit{otherwise} \end{cases}, \tag{4.15}$$

where x_i denotes the i-th miRNA or mRNA, and H_m is the m-th learned model. $\delta(x_i, H_m)$ is an indicator function and it returns one when xi appears at least once in H_m, otherwise zero. The proposed method is compared to randomly generated hypergraphs each comprising 200 hyperedges involving three miRNAs and five mRNAs. The results are derived from 100 models learned by 10 experiments of 10-fold cross validations, and 100 randomly generated hypergraphs. According to Figure 4.8 (a), our method extracts significant miRNAs only, while almost all of the miRNAs are involved in random graphs. Moreover, whereas the learning method selects several significant mRNAs, all mRNAs appear at low frequency in the random graphs, as shown to Figure 4.8 (b). The stability and reproducibility of the proposed model is evident from the high-frequency occurrence of high ranked miRNAs and mRNAs, indicating that certain genes persist in the models.

4.4.4 Constructed Higher-Order miRNA-mRNA Interaction Networks in Prostate Cancer

The miRNA-mRNA interaction network constructed from the proposed model is illustrated in Figure 4.9 and 4.10 for primary and metastatic prostate cancer respectively (Smoot et al., 2011). The constructed interaction networks comprise putative miRNA-mRNA modules associated with each stage of prostate cancer, and reflect their higher-order relationships. The primary prostate cancer network includes 67 miRNAs and 233 mRNAs, while the metastatic prostate cancer network involves 65 miRNAs and 180 mRNAs.

Many of the miRNAs in the constructed networks have been significantly asso-



(a) Appearances of miRNAs

(b) Appearances of mRNAs

Figure 4.8: Reproducibility of decisive miRNAs (a) and mRNAs (b) influencing on classification. 100 hypergraphs are generated by randomly selecting miRNAs and genes, while another 100 hypergraphs are generated by our learning method (10 experiments with 10-fold cross validation). Each hypergraph includes 200 hyperedges consisting of three miRNAs and five mRNAs. The x-axis denotes the rank of the appearance of miRNAs or mRNAs, and y-axis is the number of miRNA or mRNA appearances. Both axes are log-scaled.

ciated with prostate cancer in the literature, and are thus termed prostate cancer-related miRNAs (Jiang et al., 2009). In addition, many of the genes in the constructed networks overlap with cancer-related genes, including transcription factors. To confirm this finding, we compiled a list of 496 oncogenes and 874 tumor suppressor genes from the Cancer Genes of Memorial Sloan-Kettering Cancer Center (Higgins et al., 2007) and 1476 human transcription factors (Zhang et al., 2012a). We investigated cancer gene enrichment in the constructed interaction networks by hypergeometric test. As shown in Figure 4.11, most of the significant genes (p-value close to 0) in the constructed networks are overrepresented in the compiled list. This

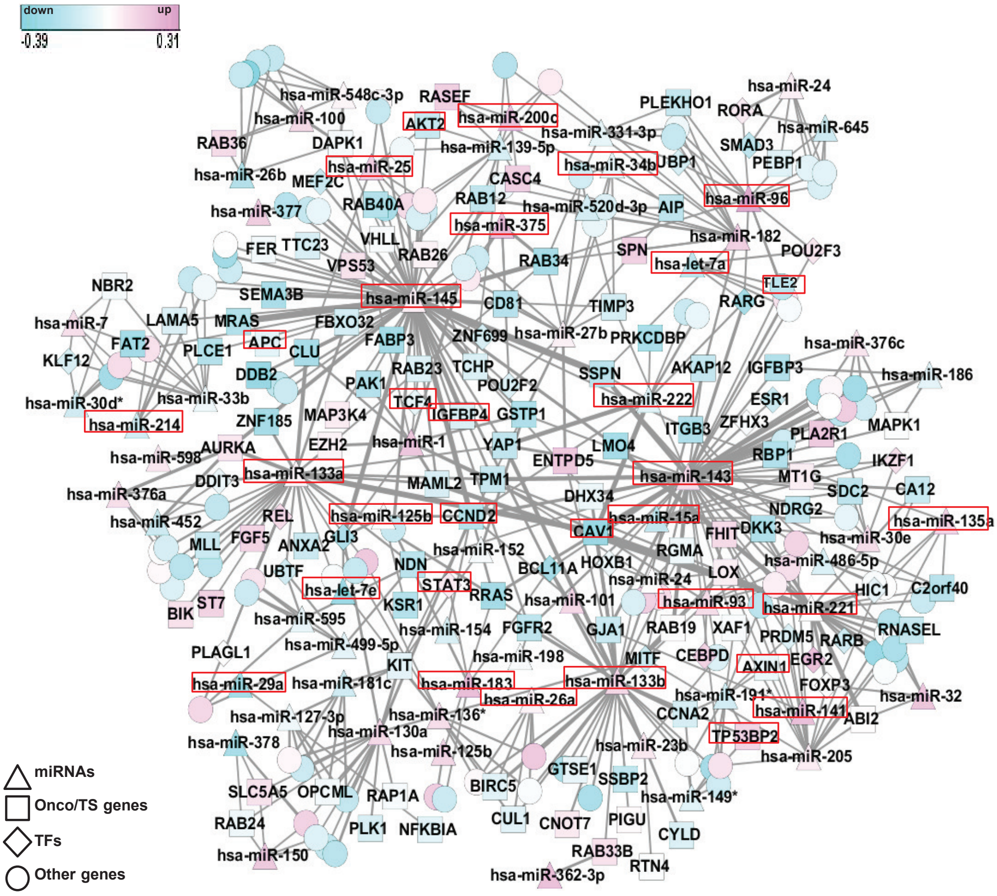


Figure 4.9: Constructed primary prostate cancer-specific miRNA-mRNA interaction networks. The primary-specific network includes 67 miRNAs and 233 mRNAs. The constructed network contains 500 bi-relational edges which are selected based on their summed weight (among all edges converted from 20000 hyperedges of 100 hypergraphs). Up- and down-expressed miRNAs and genes are determined by the mean of each stage class. The red boxed miRNAs and genes have been reported to be associated with the particular stage of prostate cancer. The triangles, rectangles, diamonds and circles denote miRNAs, oncogenes or tumor suppressor genes, transcription factors, and other genes in the network, respectively.

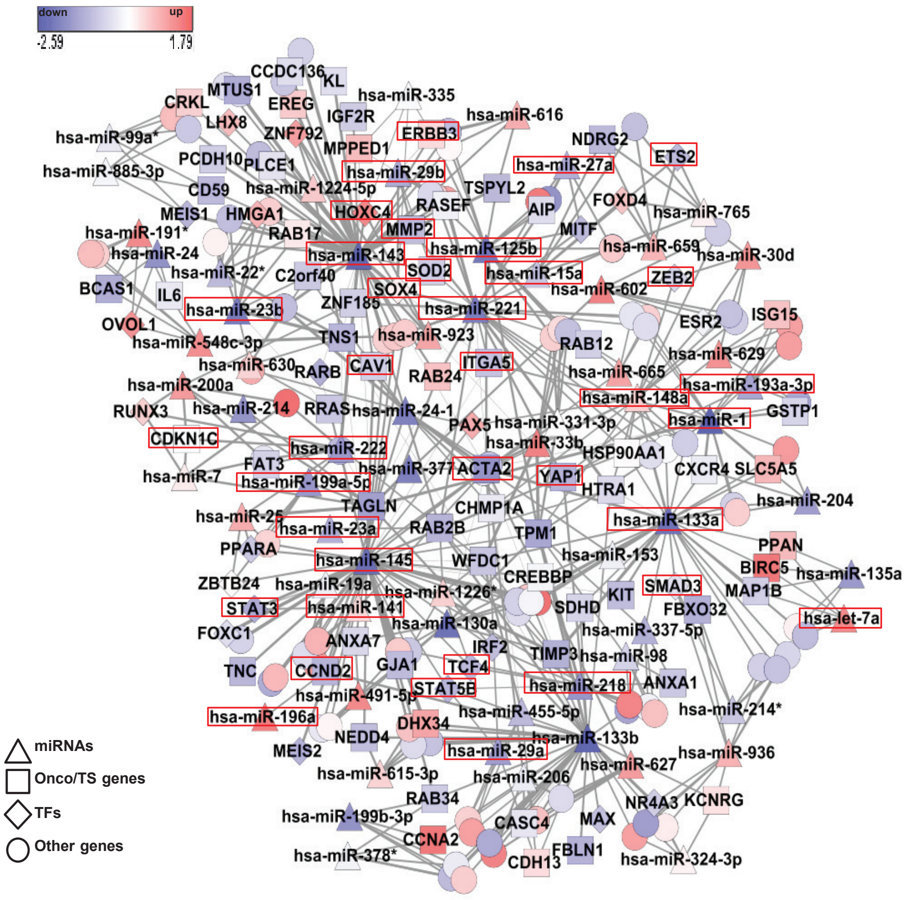


Figure 4.10: Constructed metastatic prostate cancer-specific miRNA-mRNA interaction networks. The metastatic network involves 65 miRNAs and 180 mRNAs. The constructed network includes 500 bi-relational edges which are selected based on their summed weight (among all edges converted from 20000 hyperedges of 100 hypergraphs). Up- and down-expressed miRNAs and genes are determined by the mean of each stage class. The red boxed miRNAs and genes have been reported to be associated with the particular stage of prostate cancer. The triangles, rectangles, diamonds and circles denote miRNAs, oncogenes or tumor suppressor genes, transcription factors, and other genes in the network, respectively.

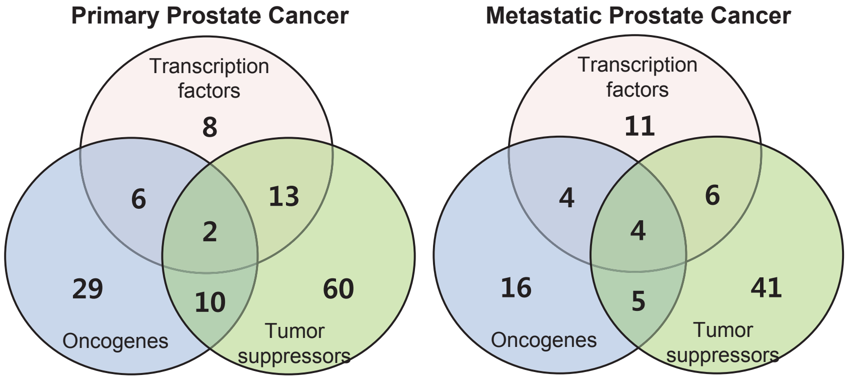
result unambiguously demonstrates that our model can build interaction networks of genetic factors associated with cancer processes.

Interestingly, the enriched hyperedges, and the expression levels of the miRNAs and mRNAs, differ considerably between the primary and metastatic networks. Up- and down-expressed miRNAs and genes are determined by their means at each stage. The red boxed miRNAs and genes are known to be associated with the various stages of prostate cancer (Coppola et al., 2010; Schaefer et al., 2010; Watahiki et al., 2011; Dasgupta et al., 2012; Gordanpour et al., 2012; Triulzi et al., 2013). The triangles rectangles, diamonds and circles denote miRNAs, oncogenes/ tumor suppressor genes, transcription factors, and other genes in the network, respectively.

4.4.5 Functional Analysis of the Constructed Interaction Networks

The constructed miRNA-mRNA interaction networks were validated by functional analyses based on a literature review and gene set analysis. As mentioned above, many of the miRNAs and mRNAs involved in the identified interactions are known indicators of prostate cancer (Coppola et al., 2010; Gordanpour et al., 2012; Watahiki et al., 2011; Schaefer et al., 2010). In addition, the mRNAs comprise a portion of their predicted target genes (Betel et al., 2010), some of which have been experimentally validated. In particular, several miRNAs are known as ′oncomiRs′ which function as oncogenes or tumor suppressors, including has-miR-1, -133a, -143, -145, -221, and -222 (Esquela-Kerscher and Slack, 2006; Kojima et al., 2011; Peng et al., 2011; Galardi et al., 2007). Many hyperedges in the constructed networks contain the above miRNAs as their components; these particular miRNAs also act as hubs in the networks.

Especially, hsa-miR-143 and hsa-miR-145 play a crucial role in metastatic prostate



Category (# total genes)	# genes in the network	*p*-value
Primary Prostate cancer		
miRNAs (96)	**28/96**	**4.06e-4**
Transcription factors (1476)	**29/1476**	**2.41e-3**
Oncogenes (495)	**47/495**	**< 0.00e-6**
Tumor suppressor genes (873)	**85/873**	**< 0.00e-6**
Metastatic Prostate cancer		
miRNAs(96)	**23/96**	**1.92e-2**
Transcription factors (1476)	**25/1476**	**8.83e-4**
Oncogenes (495)	**29/495**	**2.22e-16**
Tumor suppressor genes (873)	**56/873**	**< 0.00e-6**

Figure 4.11: The miRNAs and mRNAs in the constructed networks are enriched in cancer-related genes with a significant *p*-value

cancer, and are recognized as a clinicopathological signature of prostate cancer (Peng et al., 2011). Interaction modules involving hsa-miR-143 and -145 occupy a large portion of the networks constructed by our model. In addition, the identified interactions in metastatic prostate cancer contain several experimentally confirmed targets of hsa-miR-143 and -145, including CLINT1, CDKN1A, IRS1, MAPK7, PPM1D and SOD2. Furthermore, hsa-miR-143 and -145 are expressed at low levels in the metastatic network, as has been experimentally validated (Watahiki et al., 2011). Moreover, hsa-miR-200c emerges as a distinct miRNA in the network of primary prostate cancer. According to several studies, hsa-miR-200c overexpression inhibits metastasis prostate cancer, while aberrant regulation triggers the invasion and migration of prostate cancer at the post-transcriptional level (Vrba et al., 2010).

Our model identified several transcription factors associated with prostate cancer metastasis, such as ETS2, HOXC4, STAT3, STAT5B, SOX4 and ZEB2. Among these, SOX4, STAT3 and STAT5B are known regulators of metastatic prostate cancer through the regulation of genes involved in miRNA processing, transcriptional regulation, and developmental pathways (Scharer et al., 2009; Abdulghani et al., 2008; Gu et al., 2010). Indeed, SOX4 is directly regulated by hsa-miR-335 in cancer progression (Scharer et al., 2009), while hsa-miR-125b coordinates STAT3 regulation in the proliferation of tumor cells (Abdulghani et al., 2008).

Interactions involving hsa-miR-29b/MMP2 and hsa-miR-335/SOX4 appear concurrently in the constructed metastatic network (Table 4.3 and 4.4). This finding is consistent with previous studies, in which-miR-29b and -335 were found to suppress tumor metastasis and migration by regulating MMP2 and SOX4, respectively (Triulzi et al., 2013; Steele et al., 2010). Interestingly, both of these interactions involve hsa-miR-143, which is closely linked to prostate cancer progression. Furthermore, the well-known cancer-associated genetic factors MMP2 and SOX4 co-emerged in the identified interactions. Although the interactions identified by our model have not been previously reported, they clearly reflect higher-order relationships between miRNAs and mRNAs. As such, they may signify unknown regulatory circuits in prostate cancer development and progression. This result suggests the utility of the proposed model in identifying undiscovered miRNA-mRNA interactions.

To confirm the biological relevance of the constructed interaction networks, we analyzed the functional correlations among the network genes by canonical path-

Table 4.3: Examples of miRNA-mRNA modules (hyperedges) in primary prostate cancer

♯	miRNA and mRNA modules
1	[miR-330, miR-133$b^{1,2}$, miR-222^{1,3}, MAP1B, WWC3, $CAV1^{6}$, DHX35, TSHZ3]
2	[miR-143^{1,4}, miR-502, miR-548c, ZZEF1, C20orf194, TSPYL2, MBD3, GPR132]
3	[miR-19a^{1}, miR-133$a^{1,2}$, miR-153, BMPR1B, WWC3, PCBP4, TCEAL4, CUL4A]
4	[miR-130a, miR-375, miR-19a^{1}, RAP1A, SNORA71D, CYLD, NDUFA6, RGS9BP]
5	[miR-222^{1,3}, miR-106b, miR-222^{1,3}, ARSJ, SSPN, C3orf58, PTGDS, RARB]
6	[miR-130a, miR-133$a^{1,2}$, miR-19a^{1}, VNN1, FGF5, ELOVL7, PHPT1, RND3]
7	[miR-133$a^{1,2}$, miR-222^{1,3}, miR-130a, SCRIB, FAM108C1, EDRF1, CAR, MOXD1]
8	[miR-130a, miR-149*, miR-26a, RASEF, TPM1, CRB2, GBP, LIX1L]
9	[miR-133$b^{1,2}$, miR-23b, miR-106b, PFAS, UNC5C, HLF, PSEN1, EZH2]
10	[miR-145^{1,4}, miR-200c^{5}, miR-23b, TTC23, PARM1, TOPORS, NEBL, RCAN2]

The underlined genes are the cancer genes archived in the Memorial Sloan-Kettering Cancer Center (MSKCC)[7]. In addition, genes with a superscript number are confirmed to be related to cancer by the following literature:[1] Esquela-Kerscher and Slack, 2006, [2] Kojima et al., 2012, [3] Galardi et al., 2007, [4] Peng et al., 2011, [5] Vrba et al., 2010, [6] Kypta et al., 2012 and [7] Higgins et al., 2012.

way analysis (Liberzon et al., 2011). The significant (low p-value) results of the analysis for the primary and metastatic prostate cancer networks are summarized in Table 4.5 and 4.6. Many of the enriched pathways are closely associated with prostate tumorigenesis and metastasis. In particular, the β-catenin degradation pathway, the Wnt/β-catenin pathway and the Wnt canonical pathway are associated with Wnt signaling, which regulates many genes implicated in prostate cancer. These pathways were identified as significant in the primary prostate cancer network. Deregulation of the Wnt-related pathway reportedly affects prostate cell

Table 4.4: Examples of modules (hyperedges) in metastatic prostate cancer

♯	miRNA and mRNA modules
1	[miR-221[1,2], miR-29*b*[3], miR-143[1,4,5], *SOX4*[6,8], *MMP2*[3], RASEF, SOD2, SCN9A]
2	[miR-29*b*[3], miR-335[6], miR-143[1,4,5], *SOX4*[6,8], MPPED1, *ERBB3*[9], HOXC4, SMTN]
3	[miR-143[1,4,5], miR-22*, miR-23b, CDKN1A, HMGA1, PELO, RAB17, TMEM150]
4	[miR-125b, miR-616, miR-143[1,4,5], TSPYL2, *ERBB3*[9], ACAD8, PHF15, TMEM16G]
5	[miR-19a, miR-141, miR-145[1,4,5], PCDH20, DNAJC3, *STAT3*[10,11], ZNF385, ACTA2]
6	[miR-133*b*[1,7], miR-145[1,4,5], miR-218, IRF2, *TCF4*[12], *STAT5B*[13], RAB2B, WFDC1]
7	[miR-143[1,4,5], miR-145[1,4,5], miR-222[1,2], ITGA5, MAPK7, MAP3K2, RAB34, S100A1]
8	[miR-214, miR-143[1,4,5], miR-145[1,4,5], FEM1A, ITGA5, NAGPA, C1orf142, ERAS]
9	[miR-193b, miR-143[1,4,5], miR-145[1,4,5], CLINT1, GJA1, MAPK7, RARRES2, IL28A]
10	[miR-221[1,2], miR-1[1,7], miR-133*b*[1,7], *TPM1*[12], NDFIP2, RAD17, VPS28, INPPd5E]

The underlined genes are the cancer genes archived in the Memorial Sloan-Kettering Cancer Center (MSKCC)[14]. In addition, genes with a superscript number are confirmed to be related to cancer by the following literature:[1] Esquela-Kerscher and Slack, 2006, [2] Galardi et al., 2007, [3] Steele et al., 2010, [4] Watahiki et al., 2011, [5] Peng et al., 2011, [6] Triulzi et al., 2013, [7] Kojima et al., 2012, [8] Scharer et al., 2009, [9] Schwaetz et al., 1999, [10] Abdulghani et al., 2008, [11] Haghikia et al., 2012, [12] Kypta et al., 2012, [13] Gu et al., 2010, and [14] Higgins et al., 2012.

proliferation and differentiation (Kypta and Waxman, 2012). Moreover, the annotated genes in the constructed network, such as APC, AXIN1, AKT2, CCND2, CAV1, TLE2 and TCF4, are essential regulatory components of these pathways in prostate cancer. ErbB-related pathways were identified in the metastatic network, including the ErbB network pathway, ErbB4 pathway, Her2 pathway, ErbB2/ErbB3 signaling pathway and the EGFR pathway, which are implicated in prostate cancer progression and metastasis (Dasgupta et al., 2012; Schwartz et al., 1999). The

Table 4.5: Canonical pathway analysis of the constructed interaction networks in primary prostate cancer

Canonical Pathway Analysis	p-value (< 0.05)
Pathways in cancer	1.70e-03
Rb1 pathway	5.95e-03
Retinoic acid pathway	6.61e-03
Aurora A pathway	7.44e-03
Beta-catenin degradation pathway	9.95e-03
Wnt/beta-catenin pathway	1.03e-02
Wnt canonical signaling pathway	1.34e-02
Met pathway (signaling of HGF receptor)	1.39e-02
P38-alpha/beta downstream pathway	1.52e-02
Beta-catenin nuclear pathway	1.58e-02
Aurora B pathway	1.66e-02
EPHB forward pathway	1.81e-02
IFN-gamma pathway	1.81e-02
P53 hypoxia pathway	1.97e-02
MYC repress pathway	2.15e-02
Progesterone mediated oocyte maturation	2.19e-02
Rac CycD pathway (Ras and Rho protein on G1/S transition)	2.73e-02
PLK1 pathway	2.88e-02
IL-6 (interleukin-6) pathway	3.08e-02
FGFR2C ligand binding and activation	3.58e-02
Cell cycle	4.43e-02
PDGFR-beta signaling pathway	4.59e-02

Table 4.6: Canonical pathway analysis of the constructed interaction networks in metastatic prostate cancer

Canonical Pathway Analysis	*p*-value (< 0.05)
MYC activate pathway	1.41e-04
ErbB network pathway	2.78e-03
KIT receptor signaling pathway	3.28e-03
IL-10 pathway	4.40e-03
Pathways in cancer	4.76e-03
ErbB4 pathway	6.12e-03
Her2 pathway (ErbB2 in signal transduction and oncology)	8.51e-03
Yap1 and Wwtr1/Taz stimulated gene expression	1.09e-02
Smooth Muscle Contraction	1.22e-02
Barrestin pathway	1.53e-02
IL-6 signaling pathway	1.85e-02
STAT3 pathway	1.85e-02
IL-2/STAT5 pathway	2.00e-02
RAS pathway	2.00e-02
ErbB2/ErbB3 signaling pathway	2.19e-02
Syndecan4 pathway	2.38e-02
PPAR-alpha pathway	2.61e-02
Integrin signaling pathway	3.72e-02
Rela pathway	3.78e-02
HDAC class I pathway	3.94e-02
FOXM1 pathway	4.24e-02
IL-7 pathway	4.23e-02
EGFR pathway	4.70e-02

FOXM1 pathway also regulates tumor metastasis (including that of prostate cancer) by stimulating the expression of several genes involved in the proliferation of tumor cells and cell cycle progression (Raychaudhuri and Park, 2011). The top-ranked pathway in the metastatic network is the MYC activation pathway. MYC reportedly promotes the metastatic phenotype by altering the epigenetic landscape of cancer cells, and is overexpressed in ~75% of advanced prostate cancer patients (Dasgupta et al., 2012). Thus, the MYC pathway is a putative key feature of metastatic progression (Wolfer and Ramaswamy, 2011).

4.5 Summary

The proposed hypergraph-based model characterizes higher-order interactions among heterogeneous genetic factors from archived data. Human cancers are typically caused by the modular control of multiple genetic factors. By analyzing gene relationships at higher-order levels, thus, we can better understand the behavior of complex cancer mechanisms. Moreover, the cooperative activities and the combinatorial regulations governed by miRNAs and mRNAs are largely unknown. We have demonstrated that higher-order relationships discriminate between specific cancer stages more precisely than pair-wise analyzes of single miRNA and mRNA interactions. From this viewpoint, we can construct a more complete interaction network consisting of putative biologically significant miRNA-mRNA modules.

In addition, our method focuses on discovering potential interactions in unknown miRNA-mRNA regulatory circuits related to specific cancer stages without the known biological information (Friedman, 2004; Ivan et al., 2008). The proposed model finds statistically significant gene modules from given expression profiles using a data-driven approach with co-regulatory measure (mutual information). However, a similar hypergraph structure could be readily constructed from other types of quantitative biological information, such as miRNA-target information and gene sequence similarity values. Furthermore, the hypergraph-based model more flexibly represents miRNA-RNA interactions than other methods (which assume that the expression states of miRNAs and mRNAs are linearly proportional to each other), because it isolates significant modules from the statistical co-expressed pattern among genes at a higher-order level.

The proposed hypergraph model is similar to Bonnet's et al. (Bonnet et al., 2010a,b) and Li et al. (Liu et al., 2010), where higher-order relationships governed by miRNA-mRNA interactions are inferred solely from expression profiles. Bonnet's method is based on a clustering approach, it cannot readily infer gene regulatory modules at a specific cancer stage. In contrast to Bonnet's method, our method explicitly considers the sample status, (the primary or metastatic state of prostate cancer), from which it constructs cancer stage-specific networks. Liu's approach is based on Corr-LDA, which requires that data are discretized. By contrast, our method uses intact real-valued data, thus preventing the information loss caused by the discretization.

In brief, we have proposed a hypergraph-based model consisting of higher-order miRNA-mRNA modules, which allows the construction of biologically meaningful interaction networks associated with specific cancer stages. For identifying potential significant interactions and refining model performance, we introduced a two-phase learning approach comprising structure and parameter learning. Finally, we constructed cancer stage-specific interaction networks reflecting higher-order miRNA and mRNA relationships by converting the hypergraph structure into an ordinary graph.

We constructed higher-order miRNA-mRNA interaction networks associated with the specific stage of prostate cancer from a matched dataset using the proposed model. The performance of the proposed model is similar to that of SVMs and superior to other classification models (outperforming them by approximately 6-10 %). More importantly, our model can construct carcinogenic miRNA-hubbed networks that characterize primary and metastatic prostate cancer. Furthermore, we demonstrated that a large proportion of the miRNAs and mRNAs identified in the constructed interaction networks are indeed involved in prostate cancer progression and development. The proposed hypergraph-based model therefore presents as an alternative method for discovering potential gene regulatory circuits. Such discoveries will greatly assist our understanding of cancer pathogenesis.

Chapter 5

Hierarchical Hypergraphs for Identifying Higher-Order Genomic Interactions in Multilevel Regulation

5.1 Overview

The importance of epigenetics has been increasingly recognized in various biological processes. Epigenetic mechanisms play important roles in controlling and maintaining normal gene expression pattern via modification or rearrangement of nucleosomes by changing the accessibility of chromatin to transcriptional regulation (Bonetta, 2008). Especially, DNA methylation is a crucial epigenetic regulation in various diseases pathogenesis including carcinogenesis (Esteller, 2007; Jones, 2012). DNA methylation typically occurs at CpG islands of promoters by DNA methyltransferase (DNMT) enzyme without DNA sequence alterations, and affects transcriptional behavior in cells such as gene silencing and activating (Laird, 2010). Aberrant DNA methylation contributes to the malignant phenotype of human cancer cells as a hallmark of tumorigenesis. While cooperating with genetic alterations, also, epigenetic regulation including DNA methylation is strongly implicated in tumor initiation, development, proliferation, and suppression (Egger et al., 2004; Jones and Baylin, 2007; Handel et al., 2010). Thus, the combinatorial analysis between epigenetic and genetic factors is necessary to understand complex cancer mechanisms at the molecular level.

Ovarian cancer is one of the most deadly gynecological malignancy in the world, caused by combinatorial effects of multiple factors (Jemal et al., 2010). Abnormal DNA methylation is a common phenomenon in ovarian cancer, and closely associated with the initiation and progression of ovarian cancer by regulating multiple genetic factors such as microRNAs (miRNAs) and mRNAs (Holschneider and Berek, 2000). Herein, the coordinated regulation of miRNAs and mRNAs involved in DNA methylation should be elucidated to systemically explore the mechanism of ovarian cancer.

Here, we propose a hierarchical hypergraph model to identify higher-order miRNA-mRNA interactions associated with the regulation of DNA methylation from TCGA data (Figure 5.1). The proposed model explicitly characterizes complex relationships among multiple genomic factors involved in the specific epigenetic regulation, from which correlated gene interactions between methylome and transcriptome in biological processes including cancer pathogenesis may be identified. A hierarchy is introduced into the hypergraph model by defining two layers representing each epigenetic and genetic regulation level. The first layer consists of hyperedges that encode higher-order relationships among many genomic factors same as the traditional hypergraphs. And the second layer is composed of vari-

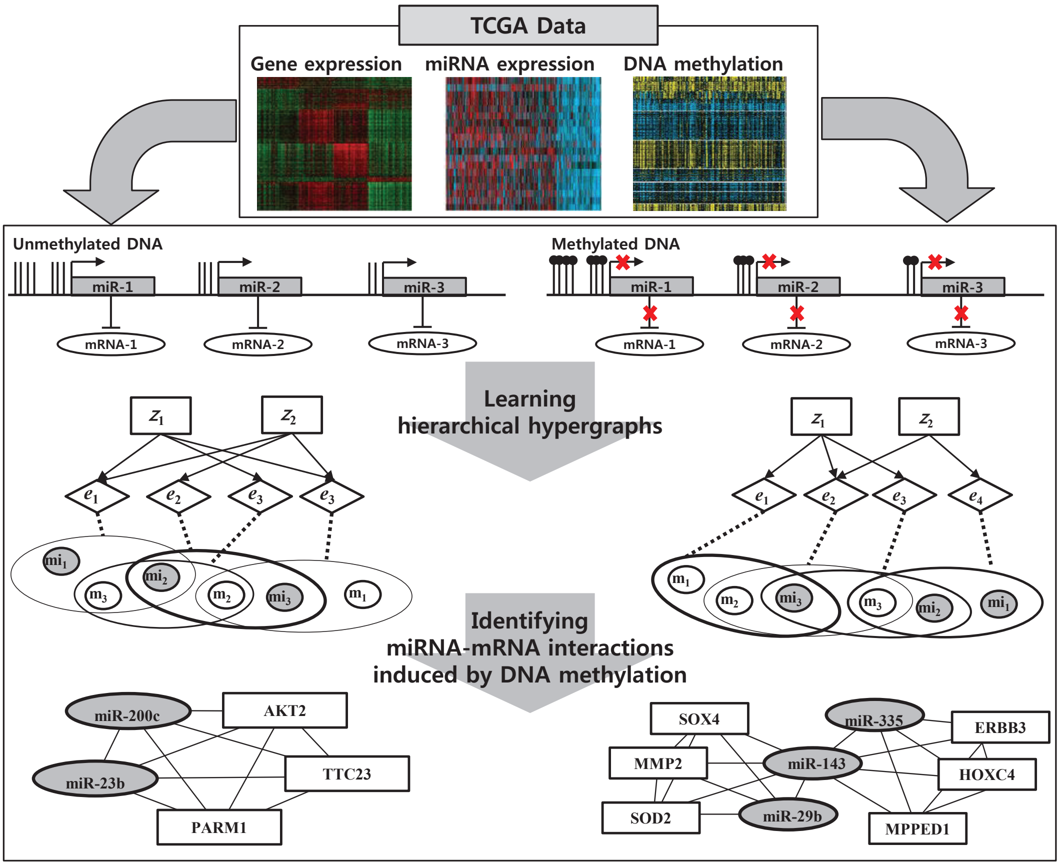


Figure 5.1: Overview of the hierarchical hypergraph for identifying higher-order genomic interactions induced by the specific DNA methylation regulation.

ables characterizing biological function and regulation. The learning of hierarchical hypergraphs proceeds by repeating three steps: generating hyperedges, calculating the objective function, and removing hyperedges with low weight. This learning method is designed upon a standard evolutionary computation framework.

The goal of the learning is to identify significant DNA methylation changes underlying cancer, and miRNA-mRNA regulatory interactions induced by the methylation change. For achieving this goal, we define an objective function which reflects the strength of interactions between miRNAs and mRNAs associated with the specific DNA methylation events from multisource genomic data using information theoretic co-regulatory measure, called mutual information. Moreover, the higher-order relationships among genomic variables are intractably complex, we adopt an evolutionary strategy for efficient searching. This hierarchical structure learning allows the model to detect potential gene regulatory circuits across the level of epigenetic, transcriptional and post-transcriptional regulation.

We identify higher-order miRNA-mRNA interactions involved in specific DNA methylation changes in ovarian cancer using TCGA data (Bell et al., 2011) from the model. We demonstrate that the proposed model can find several biologically significant miRNA-mRNA interactions implicated in DNA methylation regulation, including potential modules associated with ovarian cancer. Moreover, cancer-related miRNAs and genes dominate the identified interactions. We also confirm the biological significance of the identified interactions through literature review and functional analysis.

5.2 Analyzing Epigenetic and Genetic Interactions from Multiple Genomic Data

Recent epigenetic research has progressed to obtain a global view of gene regulation at multi-cellular level. Many studies have focused on analyzing the relationships between only two data sources, such as DNA methylation-genes (Siegfried and Simon, 2010; Spisák et al., 2012; van Eijk et al., 2012; Busche et al., 2013; Marx et al., 2013) and DNA methylation-miRNAs (Han et al., 2007; Lujambio et al., 2008;

Yan et al., 2011; Baer et al., 2012; Wong et al., 2012), on genome-wide scale from high-throughput data. These approaches have systemically investigated the complex mechanism of various cancers at the multi-level regulation. However, it is difficult to directly extract the regulatory modules between epigenetic and genetic components underlying specific cancer types. To overcome this issue, Joung et al. (Joung et al., 2013) proposed a method to extract the correlated gene pairs to DNA methylation from both expression profiles. This method calculates the unified score, consisting of the differential score and correlated score, which measures the strength of the regulatory relationships between two genes. However, the score reflects the pairwise relations of only two genes, and thus this method is difficult to precisely address complex genetic interactions associated with the epigenetic events. In recent, moreover, several studies have attempted to simultaneously explore the coordinated relationships from heterogeneous data, such as DNA methylation, gene and miRNA expression profiles, via computational approaches including statistical method (Zhu et al., 2011), matrix factorization (Zhang et al., 2012b) and regression model (Li et al., 2012). However, analyzing higher-order relationships across the levels of epigenetic, transcriptional, and post-transcriptional regulations is still rendered as a challenging issue due to the complexity of their interactions.

5.3 Hierarchical Hypergraphs for Identifying Epigenetic and Genetic Interactions

5.3.1 Hierarchical Hypergraphs

Hierarchical hypergraphs is a hypergraph-based model consisting of two distinct layers. The first layer includes hyperedges whose nodes are observable target variables while latent or observable causal variables exist in the second layer. The

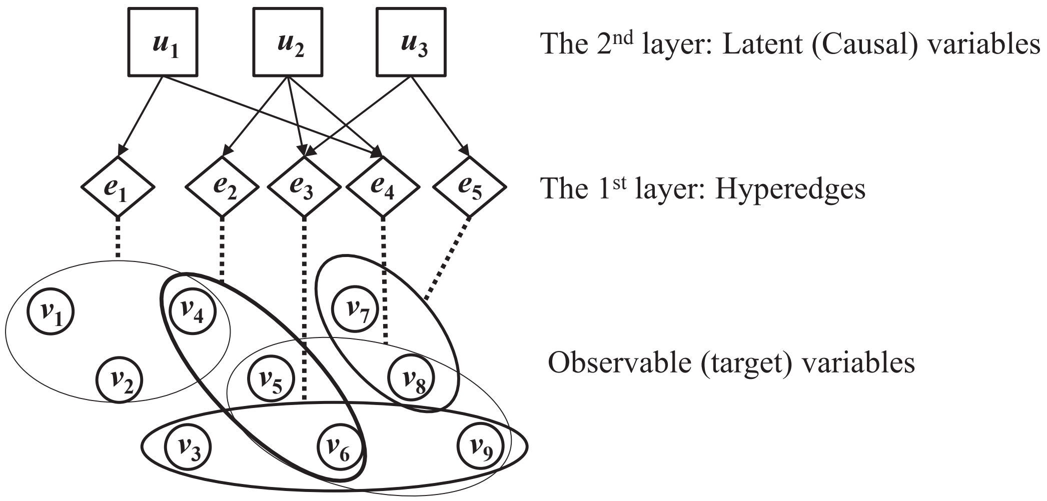

u_1
u_2
u_3
The 2nd layer: Latent (Causal) variables
e_1
e_2
e_3
e_4
e_5
The 1st layer: Hyperedges
v_1
v_4
v_7
v_2
v_5
v_8
Observable (target) variables
v_3
v_6
v_9

Figure 5.2: Graphical representation of a hierarchical hypergraph

probability distribution of observable target variables in the first layer is derived from the variables in the second layer. The probability of the latent variables in the second layer is inferred from the probability of the variables in hypergraphs in the first layer. When the second layer variables are causal variables given from the data, it is equivalent to a mixture of hypergraph classifiers with multiple class labels. A hierarchical hypergraph is similar to the model structure of deep networks. While deep networks are fully connected across the layer, the hierarchical hypergraphs are partially connected same as the previous hypergraph-based models. This partial connection allows the proposed model to be used as a soft clustering method. Moreover, the weight of our model reflects the strength of the association.

Formally, a hierarchical hypergraph H is defined as $H = (V, U, E, W)$, where V, U, E, and W are the set of vertices, causal vertices, hyperedges, and weights, respectively. Note that there exist two distinct vertex sets for observable and latent variables unlike the previous hypergraph-based models. Figure 5.2 presents the graphical representation of a hierarchical hypergraph.

For modeling the influence of DNA methylation on the expression of miRNAs and mRNAs in ovarian cancer, in this study, we set DNA methylation data attributes and miRNA and gene expression values to the causal variables in the second layer and the target variables in the first layer, respectively. Moreover, we assume three conditions for modeling miRNA-mRNA interactions by the DNA methylation events as follows:

1. DNA methylation changes influence on miRNA and mRNA expression.
2. DNA methylation events occur independently.
3. miRNAs down-regulate mRNAs and we do not consider up-regulation.

Then, each hyperedge represents the higher-order combinatorial modules of miRNAs and mRNAs, and a DNA methylation change is associated with a group of several hyperedges, and hyperedges involved in the same group show the similar expression pattern each other with respect to the DNA methylation. By introducing explicit causal variables into the model, therefore, this hierarchical structure allows the hypergraph-based model to clearly characterize miRNA and mRNA expression patterns induced by a DNA methylation change. When multisource expression profiles are given as a dataset $D = \{d^{(n)}\}_{n=1}^{N} = \{(\mathbf{x},\mathbf{y},\mathbf{z})^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ denote the vector of miRNA, mRNA, and DNA methylation expression variables, the hierarchical hypergraphs are represented like Figure 5.3. In figure 5.3, a hyperedge in the first layer represents a higher-order relationship among more than two miRNAs and mRNAs. A DNA methylation controls several hyperedges whose miRNAs and mRNAs are influenced by the methylation. There is no interconnection between nodes in the second layer by the assumption 1.

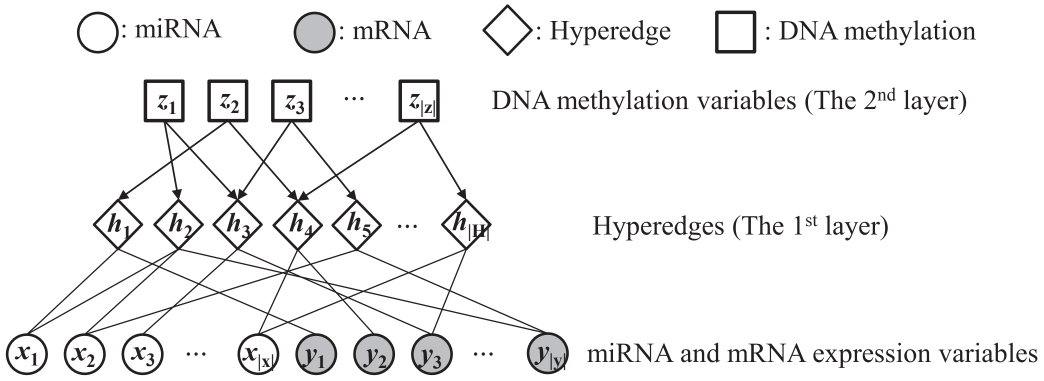

: miRNA
: mRNA
: Hyperedge
: DNA methylation
z_1 z_2 z_3 ⋯ $z_{|\mathbf{z}|}$
DNA methylation variables (The 2nd layer)
h_1 h_2 h_3 h_4 h_5 ⋯ $h_{|\mathbf{H}|}$
Hyperedges (The 1st layer)
x_1 x_2 x_3 ⋯ $x_{|\mathbf{x}|}$ y_1 y_2 y_3 ⋯ $y_{|\mathbf{y}|}$
miRNA and mRNA expression variables

Figure 5.3: Hierarchical hypergraph model

5.3.2 Learning Hierarchical Hypergraphs

The learning of hierarchical hypergraphs proceeds by repeating three steps: generating hyperedges, calculating the objective function, and removing hyperedges with low weight. This learning procedure resembles a conventional evolutionary framework including variation, fitness computation, and selection, such as genetic algorithm and genetic programming.

Fitness Computation

The goal of the learning is to identify significant DNA methylation sites under a specific disease and miRNA-mRNA regulatory modules induced by the methylation change from the multisource data. For achieving this goal, we define an objective function which reflects the strength of relationships among DNA methylation and miRNA-mRNA expression level. In specific, the objective function of the model at time t $f(H_t)$ is calculated by summing the function computing the interaction strength between a hyperedge e and a methylation $\mathbf{z}$ $g(e; z)$:

$$f(H_t) = \sum_{z \in \mathbf{z}} \sum_{e \in E_t^z} g(e; z) = \sum_{z \in \mathbf{z}} \sum_{e \in E_t^z} I(e; z), \tag{5.1}$$

where $g(e; z)$ is a interaction function and E_t^z denotes a subset of hyperedges strongly associated with $\mathbf{z}$ of H_t. We use multivariate mutual information (Kraskov et al., 2004) as the interaction function in this study. Mutual information (MI) is an information-theoretic measure defined as the difference between two entropy and it reflects the degree of conditional independency between two random variables:

$$\begin{aligned} I(A; B) &= H(A) - H(A|B) \\ &= \sum_{a \in A} \sum_{b \in B} p(a, b) \log\left(\frac{p(a,b)}{p(a)p(b)}\right), \end{aligned} \tag{5.2}$$

where A and B are random variables. When a methylation has stronger influences on the expression of a genetic factor, therefore, MI between the gene and the methylation goes higher. Multivariate MI (MMI) is a generalized measure where the number of variables is extended to three or more variables:

$$\begin{aligned} I(e; z) &= I(x^e; y^e; z) = I(x^e; y^e) - I(x^e; y^e|z) \\ &= I(x^e; y^e) - E_z\left(I(x^e; y^e)|z\right), \end{aligned} \tag{5.3}$$

where x^e and y^e denote miRNA and mRNA vertices involved in a hyperedge e.

Variation: Hyperedges Generation

The main issue of the learning is to generate hyperedges consisting of miRNAs and mRNAs strongly related the methylations and this involves searching the huge combinatorial feature space due to the definition of hyperedge. Especially, it is infeasible to search all the space because the number of variables is very large in biological data. For efficient searching, we use an evolutionary method based on correlation coefficients between genetic variables for generating hyperedges. We assume that the genetic modules relevant to a methylation are composed of miRNA-miRNA pairs and miRNA-mRNA pairs whose change of correlation coefficients is larger depending on the methylation event. Specifically, a hyperedge is generated as follows:

1. Select the i-th DNA methylation z_i from the methylation vector $\mathbf{z}$. E_z is set to an empty set.

2. Divide miRNA-mRNA data samples into two separate groups based on the methylation threshold. We use the mean of the methylation on all samples as the threshold.

3. Make two correlation coefficient matrices from two groups including miRNA-miRNA matrix $C^i_{\mathbf{xx}}$ and miRNA-mRNA matrix $C^i_{\mathbf{xy}}$. Each matrix are calculated as follows:

$$\begin{aligned} C^i_{\mathbf{xx}}(m,n) &= \left|C^+_{\mathbf{xx}}(m,n)\right| + \left|C^-_{\mathbf{xx}}(m,n)\right|, \\ C^i_{\mathbf{xy}}(m,n) &= \begin{cases} \left|C^+_{\mathbf{xy}}(m,n) + C^-_{\mathbf{xy}}(m,n)\right|, & \text{if } C^+_{\mathbf{xx}}(m,n) + C^-_{\mathbf{xx}}(m,n) < 0 \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \tag{5.4}$$

where $C^+_{\mathbf{xx}}$, $C^-_{\mathbf{xx}}$, $C^+_{\mathbf{xy}}$ and $C^-_{\mathbf{xy}}$ denote methylated miRNA-miRNA, unmethylated miRNA-miRNA, methylated miRNA-mRNA matrix, and unmethylated miRNA-mRNA matrix for z_i, respectively. $C(m,n)$ is the element in the m-th row and the n-th column of C. This definition allows two matrices to reveal the change of the expression of genetic factors, and especially the definition of $C^i_{\mathbf{xy}}$considers the third assumption in section 2.

4. A hyperedge e is set to an empty set.

5. Select a miRNA x' from $\mathbf{x}$ based on the probability $P(x_j|z_i)$:

$$P(x' = x_j|z_i) = \frac{\Delta(x_j|z_i)}{||\Delta(\mathbf{x}|z_i)||_1}, \ \Delta(x|z_i) = \left|\bar{\mathbf{x}}_{zi}^{+} - \bar{\mathbf{x}}_{zi}^{-}\right| \tag{5.5}$$

   where $\bar{\mathbf{x}}_{zi}^{+}$ and $\bar{\mathbf{x}}_{zi}^{-}$ denote the vector of mean miRNA expression value on methylated and unmethylated cases for given z_i, respectively. Also, $||C||_1$ denotes the summation of all elements in matrix C. Then, x' is added into e.

6. Select a pair of two mRNAs including x' from $\mathbf{x}$ based on the selection probability:

$$P(x = x_k|x', z_i) = \frac{C_{x'\mathbf{x}}^{i}(k)}{\left\|C_{x'\mathbf{x}}^{i}\right\|_1}. \tag{5.6}$$

   where $C_{x'\mathbf{x}}^{i}$ is the row vector of the index of x' of $C_{\mathbf{xx}}^{i}$. After that, add the selected miRNA into e.

7. Determine a miRNA x'' in e randomly and then select a mRNA from $\mathbf{y}$ based on the selection probabilities:

$$P(y = y_m|x'', z_i) = \frac{C_{x''\mathbf{y}}^{i}(m)}{\left\|C_{x''\mathbf{y}}^{i}\right\|_1}. \tag{5.7}$$

   where $C_{x''\mathbf{y}}^{i}$ is a row vector corresponding to the index of x'' from $C_{\mathbf{xy}}^{i}$. After that, add the selected mRNA into e.

8. Repeat the steps 6) and 7) as the predefined times and the $d(e)$ is equal to the number of all the miRNAs and mRNAs involved in e.

9. Add e into E^i and repeat from step 4).

The amount of hyperedges for a methylation is given as a parameter. The above approach allows a hyperedge to consist of relevant miRNAs and mRNAs influenced by a specific DNA methylation change. Note that the third assumption in the previous section is applied to making the correlation coefficient matrices in (5.4). That is, positive values are ignored to be zero in two miRNA-mRNA matrices because we consider down-regulation only. Due to the same reason, we ignore negative values in two miRNA-miRNA matrices. Therefore, this method enhances the efficiency of the learning by being utilized as a guided searching strategy. Dissimilar to using a feature selection method, however, this method does not exclude the space represented as the miRNA-mRNA relationships completely by using probabilistic selection.

Selection: Hyperedges Replacement

Replacing hyperedges is the process for model selection while generating hyperedges is conducted for enhancing the model variation. Hyperedges with low weight are eliminated from the model and new hyperedges are generated. The structure of the model is evolved to be constructed a hypergraphs composed of significant miRNA-mRNA modules associated with the methylations via the hyperedge substitution. When removing the hyperedges, the elimination is carried out per hyperedge group. The amount of replaced hyperedges is determined by learning epochs as follows:

$$R_t = \frac{R_{max} - R_{min}}{\exp\left(t/\kappa\right)} + R_{min}\,, \tag{5.8}$$

where R_{max} and R_{min} denote the maximum and minimum boundary values of R_t,

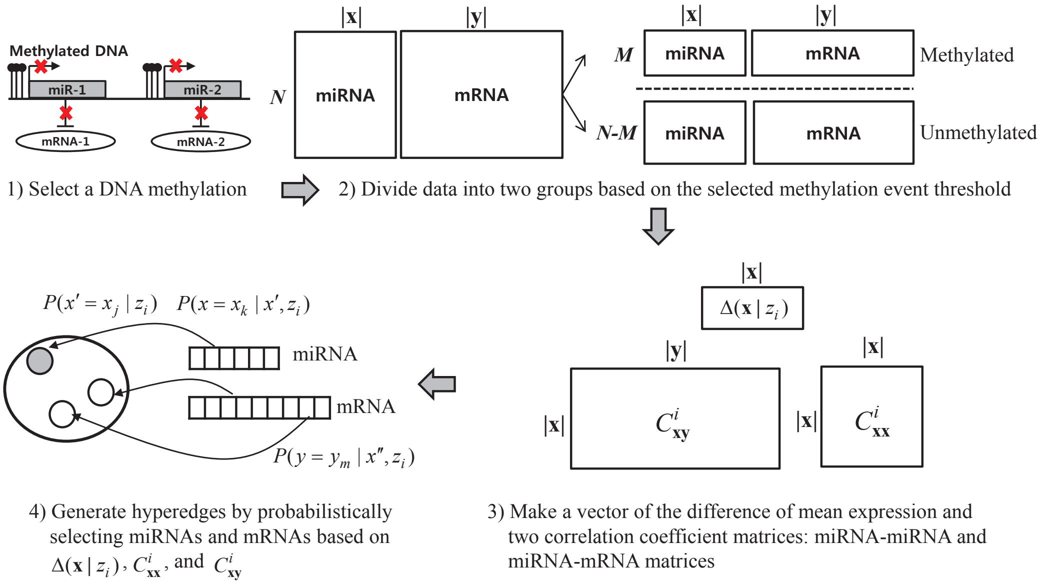

Methylated DNA
miR-1
miR-2
mRNA-1
mRNA-2
1) Select a DNA methylation
|x|
|y|
N
miRNA
mRNA
M
N-M
Methylated
Unmethylated
2) Divide data into two groups based on the selected methylation event threshold
Δ(x | z_i)
C^i_xy
C^i_xx
3) Make a vector of the difference of mean expression and two correlation coefficient matrices: miRNA-miRNA and miRNA-mRNA matrices
P(x' = x_j | z_i)
P(x = x_k | x', z_i)
P(y = y_m | x'', z_i)
4) Generate hyperedges by probabilistically selecting miRNAs and mRNAs based on Δ(x | z_i), C^i_xx, and C^i_xy

Figure 5.4: Learning procedure for hierarchical hypergraphs

respectively, and κ is a constant to moderate the speed from R_{max} to R_{min}. The learning procedure of the hierarchical hypergraph model is presented in Figure 5.4.

5.4 Identifying Higher-Order Genomic Interactions in Multilevel Regulation

5.4.1 Data and Experimental Settings

Ovarian cancer is the fifth-leading cause of cancer death among women in the United States Jemal et al. (2010). Most deaths are of patients presenting with advanced-stage, high-grade serous ovarian cancer (HGS-OvCa). Approximately 13% of of HGS-OvCa is attributable to germline mutations in BRCA1 and BRCA2 and a smaller percentage can be accounted for by other germline mutations. However, most ovarian cancer can be attributed to a growing number of somatic aberrations. For this reason, a catalogue of molecular aberrations that cause ovarian cancer is critical for developing and deploying therapies that will improve patients' lives. Thus, the identification of molecular abnormalities that influence pathophysiology in ovarian cancer needs a large-scale integrative view which can comprehensively investigate genomic and epigenomic relationships on clinically annotated HGS-OvCa samples.

In this study, the DNA methylation, miRNA and mRNA expression data in ovarian cancer were obtained from TCGA (Bell et al., 2011). The Cancer Genome Atlas project is generating multi levels of the key genomic changes such as DNA methylation, miRNA expression, and mRNA expression, for the same set of cancer samples. We use three types of data, as follows: DNA methylation data (Illumina 27K), mRNA expression data (Agilent G4502A) and miRNA expression data (Agilent H-miRNA_8x15K v2). In total, 385 ovarian cancer samples are shared by the three datasets. The dataset contains the expression profiles of 799 miRNAs and 16046 mRNAs and the DNA methylation profiles of 15418 CpG loci. During preprocessing, we normalized the columns of the expression matrices, and then scaled all the matrices so that sum of squares of each matrix is the same. The experimental parameter settings are listed in Table 5.1. The parameters are those yielding optimal performance in empirical experiments. A hypergraph can include hyperedges with different number of genetic variables but we fixed the number of variables for all hyperedges of a hypergraph in this study.

Table 5.1: Parameter settings for experiments

Parameters	Values
♯ of miRNAs, ♯ of mRNAs	2, 3
♯ of hyperedges per DNA methylation	100
R_{max}, R_{min}	0.5, 0.1
κ in (5.8)	1.0
Iteration number	100

5.4.2 Identified Higher-Order miRNA-mRNA Interactions Induced by DNA Methylation in Ovarian Cancer

Figure 5.5 present learning curves in hierarchical hypergraph learning phases. As the measure for objective function, we used a summation of multivariate mutual information of all hyperedge group on each DNA methylation event because the aim of the learning is to find the significant higher-order genomic interactions induced by the specific DNA methylation change, and MMI is the measure reflecting the strength of interactions among genomic factors in hyperedges considering the methylation event.

The proposed model has provided sets of genomic features from different regulatory layers that are likely to be synergistic in their impact on mRNA expression profiles. To further elaborate the relationships between those implicated features, we used the functional analysis to identify molecular interactions. From each the learned model, we identified a set consisting of miRNAs and mRNAs involved in the methylation-adjacent genes.

Table 5.2 presents an example of the identified miRNA-mRNA interactions asso-

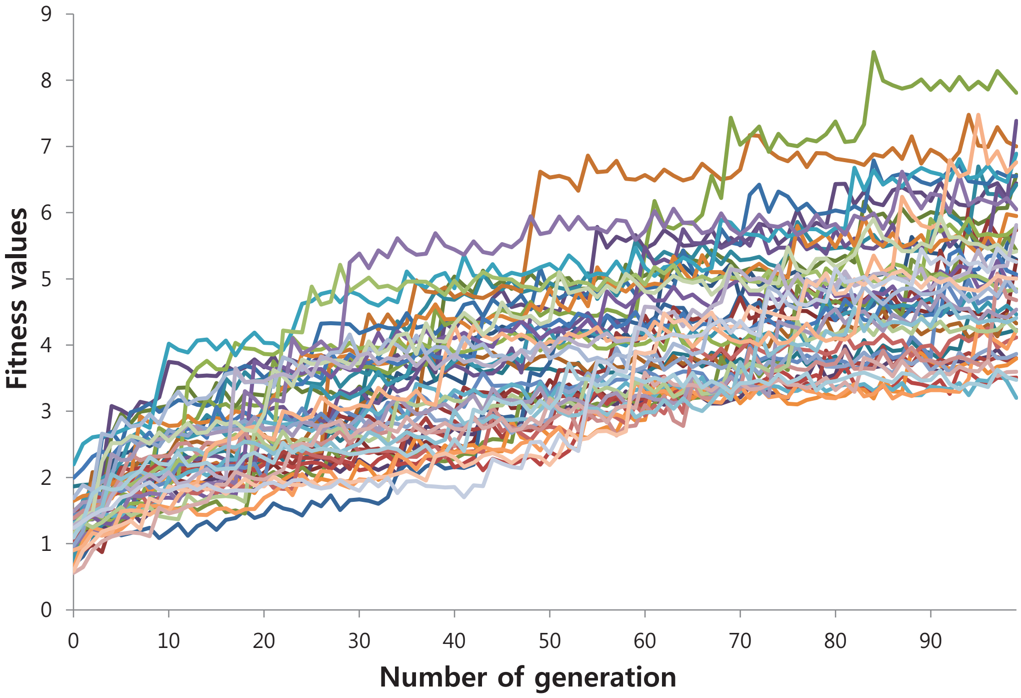


Figure 5.5: Learning curves in hierarchical hypergraph learning phases. The objective function is defined as a summation of MMI of all hyperedge group on each DNA methylation event.

ciated with the specific DNA methylation. In particular, we observe that miRNA-mRNA interactions involving DNA methylated ESR2 and THBS2 from the learned model. The estrogen receptor beta gene (ESR2) might influence epithelial ovarian risk through regulation of cell proliferation and apoptosis. Various studies have shown that methylation of the ESR2 is associated with reduced expression of ESR2 isoforms in breast, prostate and ovarian cancer tissue and cell lines (Philips et al., 2012; Pearce et al., 2008). Also, thrombospondin-2 (THBS2) is known as a regulator

Table 5.2: Identified higher-order miRNA-mRNA interactions induced by a specific DNA methylation (hyperedges) in ovarian cancer

♯	DNA Methylation	miRNA and mRNA interactions
1	LMTK2 →	[hsa-mir-200$a^{4,5}$, hsa-mir-200$b^{4,5}$, OTX2, ZEB2^{6}, SPN]
2	THBS2^{1} →	[hsa-mir-502-3p, hsa-mir-500*, FGF3, CUL2, CSF3]
3	ESR2^{2,3} →	[hsa-let-7$b^{4,7}$, hsa-mir-130$b^{4,8}$, EZH2^{9}, FAT3, MTHFR]
4	ESR2^{2,3} →	[hsa-mir-514, hsa-mir-507, RAB3D, PALB2, HUWE1]
5	THBS2^{1} →	[hsa-mir-154, hsa-mir-337-5p, MCM3, MTHFR, LIMD1]
6	FRZB →	[hsa-mir-508-3p, hsa-mir-507, ZEB2^{6}, PDCD4, FLT3]
7	THBS2^{1} →	[hsa-mir-34a^{4}, hsa-mir-197, SERPINB5, CHK1^{4}, EDN3]
8	ESR2^{2,3} →	[hsa-mir-509-3p, hsa-mir-514, USP4, MAP2K3, PATZ1]
9	LMTK2 →	[hsa-let-7$f^{4,7}$, hsa-mir-595, IL24^{4}, BCAR1, DAPK1]
10	ESR2^{2,3} →	[hsa-mir-34a^{4}, hsa-mir-377, IL6^{4}, CDH17, EXTL3]

Genes with a superscript number are confirmed to be related to cancer by the following literature:[1] Czekierdowski et al., 2008, [2] Pearce et al., 2008, [3] Philips et al., 2012, [4] Leva and Croce, 2013, [5] Gregory et al., 2008, [6] Wu et al., 2011, [7] Lu et al., 2007, [8] Yang et al., 2013, and [9] Li and Zhang, 2013

of ovarian cancer through the regulation of genes involved in transcriptional regulation and developmental pathways (Czekierdowski et al., 2008). Such interactions involving ESR2 and THBS2 appear in the learned model. This finding is consistent with previous studies.

The identified miRNA-mRNA interactions by our model contained the known indicators of ovarian cancer (Leva and Croce, 2013). EZH2 promotes cell proliferation, inhibits apoptosis and enhances angiogenesis in epithelial ovarian cancer, and its target genes are involved in a variety of biological processes such as stem cell pluripotency, cell proliferation, and oncogenic transformation (Li and Zhang, 2013). ZEB2 has important functions in metastasis of ovarian cancer. miR-200a has been reported to be a prognostic marker and to play an important role in ovarian cancer progression. Especially miR-200a down-regulates ZEB2 level, resulting in decreased ovarian cancer stem cells migration and invasion (Wu et al., 2011) and ZEB1/2, two transcription factors involved in the mediation of the epithelial to mesenchymal transition, can inhibit the expression of miR-200 family members by binding to the promoter of both miR-200 clusters thereby blocking transcription (Gregory et al., 2008). Such well-known miR-200a and ZEB2 co-merge in the identified interactions. Hsa-let-7a-3 is methylated in epithelial ovarian cancer, and low expression of let-7a is associated with poor prognosis (Lu et al., 2007). Epigenetic silencing of miR-130b through hypermethylation of the adjacent CpG island has been also identified and low expression of miR-130b was correlated to ovarian cancer with high stage and multidrug resistance (D. Yang et al., 2013). In fact, treatment of ovarian-cancer cells with demethylating agents increased miR-130b levels and decreased the IC50 of paclitaxel and cisplatin treatment.

5.5 Summary

Epigenetic and genetic abnormalities are observed in various types of cancer. In particular, DNA methylation regulation cooperates with genetic factors, such as miRNAs and mRNAs, to affect biological processes including carcinogenesis and cancer progression. The complexity of epigenetic and genetic interactions is a major barrier to identifying their co-regulatory activities and functional roles.

We have proposed hierarchical hypergraph models consisting of the observable (target) layer including genomic variables, and the latent (causal) layer comprised of epigenetic regulatory variables, which allow the identification of higher-order miRNA-mRNA interactions induced by DNA methylation changes. For identifying potential significant interactions across the multilevel regulation, we introduced a hierarchical structure into the model. Finally, we found the higher-order genomic interactions by calculating co-regulatory strength between miRNAs and mRNAs implicated in tumor-specific epigenetic events from multi source data.

We identified cancer-specific genomic interactions associated with the specific DNA methylation changes from TCGA data using the proposed model. We demonstrated that a large proportion of the miRNAs and mRNAs in the identified interactions are well known to be involved in ovarian cancer progression and development. Therefore, the proposed hierarchical hypergraph model seems to be a promising approach to extract higher-order coherent substructures from large-scale and multidimensional genomic data.

Chapter 6

Concluding Remarks

6.1 Summary of the Dissertation

In this dissertation, we propose a novel hypergraph model for analyzing diverse biological problems characterizing higher-order interactions among many genomic factors, and the improved learning method for efficiently searching huge problem spaces representing their complex relationships. In addition, the graph-analyzing method is proposed to effectively extract meaningful biological information and knowledge from the learned models.

The proposed model structure has shown potential for modeling complex biological phenomena from different types of genomic data due to its high power of representation and great flexibility. Moreover, it is possible to intuitively extract significant hyperedges in a interpretable form from the learned models which is a useful property for data mining problems in biology and medicine.

The learning of hypergraph models involves searching a huge combinatorial feature space due to its definition and the problem space exponentially enlarges as the number of features increase. For this reason, we apply an evolutionary computation to the learning method of the hypergraph model, and introduce information-theoretical criteria into the evolutionary learning mechanism for more efficiently searching the high-dimensional problem space representing higher-order relationships among many genetic variables.

Furthermore, we suggest the advanced model structure to extensively analyze biological processes from multiple genomic data. Hence, a hierarchy is introduced into the hypergraph-based model by defining two layers representing different regulatory levels such as epigenetic and genetic stage. The first layer consists of hyperedges that encode higher-order relationships among many genetic factors same as the conventional hypergraphs. And the second layer is composed of latent variables characterizing biological function and regulation. This hierarchical structure allows the proposed hypergraph model to explicitly represent gene regulatory circuits as functional blocks or groups across the multiple regulation level for better understanding complex biological processes such as human cancers.

Lastly, we propose a new method to enable to identify co-regulatory gene modules or to construct gene regulatory networks from the learned hypergraph model. It is important to extract meaningful information and knowledge from the learned model in biological and medical field. A network characterizing higher-order genomic interactions is constructed from the learned hypergraphs based on a minimum-cut approach in this dissertation. Thus, the proposed model can extract meaningful knowledge such as co-regulatory gene modules, interactions, pathways or networks from various genomic data. Furthermore, it can discover new or potential genomic regulatory circuits which assist our understanding of biological systems including cancer pathogenesis.

In this dissertation, we demonstrate the proposed hypergraph-based model explicitly represents the complex relationships such as miRNA-mRNA interactions and efficiently learns cancer-specific higher-order patterns from multiple cancer genomic profiles. The performance of the proposed model is validated on various high-dimensional datasets with tens of thousands of genetic variables. Experimental results show the proposed model is competes with or outperforms other state-of-the-art models such as naive Bayes classifier, decision tree, random forest, AdaBoost, support vector machine, and learning classifier system. The proposed learning method based on introducing information theory enhances the performance of the hypergraph model, and moreover decreases the learning time while achieving the same accuracies. More importantly, our approach can extract biologically meaningful knowledge from the learned model such as prognostic cancer gene modules, cancer stage-specific miRNA-mRNA interaction networks and carcinogenic miRNA-mRNA groups associated with specific DNA methylation events. The biological significance of the extracted genomic modules, interactions and networks is confirmed through literature review and functional analysis. Herein, the proposed hypergraph model is useful for identifying new or potential gene regulatory circuits. Moreover such discoveries will greatly assist our understanding of cancer mechanisms. Thus, our model presents as an alternative method for solving a variety of biological and medical problems.

6.2 Directions for Further Research

The present work can be extended into several directions. First of all, unlike other models, the proposed model can efficiently handle the very high-dimensional data required for complex higher-order interactions among features. However, the limitation of the proposed model emerges at small sample sizes. If the data are few, the reliability of the mean and covariance defined in a hyperedge is reduced.

Another direction of future work is the proposed hypergraph model has relatively heavy time complexity which is O(MN), where M and N denote the number of hyperedges and data instances, respectively. Although our method conducts simple operation without any complex numerical function in the learning, it spends much time for learning from data with large number of instances. This issue can be solved by using parallel processing techniques such as GPGPU. Lastly, the other is that our learning method is batch-style and thus it is not suitable for learning from increasing data. We will improve the method to allow the model to incrementally learn.

Bibliography

Abdulghani, J., Gu, L., Dagvadorj, A., et al. (2008). STAT3 promotes metastatic progression of prostate cancer. *The American Journal of Pathology*, 172(6):1717–1728.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92:291–294.

Baer, C., Claus, R., Frenzel, L., et al. (2012). Extensive promoter DNA hypermethylation and hypomethylation is associated with aberrant microRNA expression in chronic lymphocytic leukemia. *Cancer research*, 72(15):3775–3785.

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M. (2010). Development of the human cancer microrna network. *Silence*, 1(1):6.

Bar-Joseph, Z., Gerber, G., Lee, T., Rinaldi, N., Yoo, Y., Robert, F., Gordon, D., Fraenkel, E., Jaakkola, T., Young, R., and Gifford, D. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11):1337–1342.

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

Bell, D., Berchuck, A., Birrer, M., et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

Berchuck, A., Iversen, E. S., Lancaster, J. M., et al. (2005). Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin Cancer Res*, 11:3686–3696.

Berge, C. (1989). *Hypergraphs*. New York: Elsevier Science Publishers.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11(8):R90.

Bishop, C. and Nasrabadi, N. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Bonetta, L. (2008). Epigenomics: Detailed analysis. *Nature*, 454(7205):795–798.

Bonneau, R. (2008). Learning biological networks: from modules to dynamic. *Nat Chem Biol*, 4:658–664.

Bonnet, E., Michoel, T., and de Peer, Y. V. (2010a). Prediction of a gene regulatory network linked to prostate cancer from gene expression, microrna and clinical data. *Bioinformatics*, 26(18):638–644.

Bonnet, E., Tatari, M., Joshi, A., Michoel, T., Marchal, K., Berx, G., and de Peer, Y. V. (2010b). Module network inference from a cancer gene expression data set identifies microRNA regulated modules. *PLoS One*, 5(4):e10162.

Bonneville, R. and Jin, V. (2013). A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor α target genes. *Bioinformatics*, 29(1):22–28.

Boulesteix, A. L., Porzelius, C., and Daumer, M. (2008). Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698–1706.

Braitman, L. E. and Davidoff, F. (1996). Predicting clinical states in individual patients. *Ann Int Med*, 125:406–412.

Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., and He, X. (2010). Music recommendation by unified hypergraph: Combining social media information and music content. In *Proceedings of the International Conference on Multimedia*, pages 391–400.

Buhmann, M. (2003). *Radial basis functions: theory and implementations*, volume 12. Cambridge university press.

Busche, S., Ge, B., Vidal, R., et al. (2013). Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. *Cancer Research*, 73(14):4323–4336.

Chen, M., Cho, J., and Zhao, H. (2011). Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genetics*, 7(4):e1001353.

Coppola, V., Maria, R., and Bonci, D. (2010). MicroRNAs and prostate cancer. *Endocrine-Related Cancer*, 17(1):F1–F17.

Cruz-Ramírez, N., Acosta-Mesa, H., Carrillo-Calvet, H., et al. (2007). Diagnosis of breast cancer using Bayesian networks: A case study. *Computers in Biology and Medicine*, 37(11):1553–1564.

Czekierdowski, A., Czekierdowska, S., Danilos, J., et al. (2008). Microvessel density and cpg island methylation of THBS2 gene in malignant ovarian tumors. *J Physiol Pharmacol*, 59(Suppl 4):53–65.

D. Yang, D., Sun, Y., Hu, L., et al. (2013). Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell*, 23(2):186–199.

Dasgupta, S., Srinidhi, S., Vishwanatha, J., et al. (2012). Oncogenic activation in prostate cancer progression and metastasis: Molecular insights and future challenges. *Journal of Carcinogenesis*, 11(1):4.

Demirkaya, O., Asyali, M., and Shoukri, M. (2005). Segmentation of cDNA microarray spots using Markov random field modeling. *Bioinformatics*, 21(13):2994–3000.

Donk, N. v., Bloem, A., Lokhorst, H., et al. (2006). New treatment strategies for multiple myeloma by targeting BCL-2 and the mevalonate pathway. *Current Pharmaceutical Design*, 12(3):327–340.

Edakunni, N., Kovacs, T., Brown, G., and Marshall, J. (2009). Modeling UCS as a mixture of experts. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, pages 1187–1194.

Eddy, S. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3):361–365.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755–763.

Egger, G., Liang, G., Aparicio, A., and Jones, P. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463.

Esquela-Kerscher, A. and Slack, F. (2006). Oncomirs: microRNAs with a role in cancer. *Nature Reviews Cancer*, 6(4):259–269.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298.

Fan, X., Shi, L., Fang, H., Cheng, Y., Perkins, R., and Tong, W. (2010). DNA microarrays are predictive of cancer prognosis: a re-evaluation. *Clin Cancer Res*, 16:629–636.

Fogel, G. B. (2008). Computational intelligence approaches for pattern discovery in biological systems. *Brief Bioinform*, 9(4):307–316.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620.

Galardi, S., Mercatelli, N., Giorda, E., Massalini, S., Frajese, G., Ciafrè, S., and Farace, M. (2007). miR-221 and miR-222 expression affects the proliferation potential of human prostate carcinoma cell lines by targeting p27Kip1. *Journal of Biological Chemistry*, 282(32):23716–23724.

Gallo, G., Longo, G., Pallottino, S., and S, S. N. (1993). Directed hypergraphs and applications. *Discrete Appl Math*, 42(4):177–201.

Gevaert, O., Smet, F., Timmerman, D., Moreau, Y., and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):184–190.

Goodison, S., Sun, Y., and Urquidi, V. (2010). Derivation of cancer diagnostic and prognostic signatures from gene expression data. *Bioanalysis*, 2(5):855–862.

Gordanpour, A., Nam, R., Sugar, L., and Seth, A. (2012). MicroRNAs in prostate cancer: from biomarkers to molecularly-based therapeutics. *Prostate Cancer and Prostatic Diseases*, 15(4):314–319.

Gregory, P., Bert, A., Paterson, E., et al. (2008). The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature cell biology*, 10(5):593–601.

Gu, L., Vogiatzi, P., Puhr, M., et al. (2010). STAT5 promotes metastatic behavior of human prostate cancer cells in vitro and in vivo. *Endocrine-Related Cancer*, 17(2):481–493.

Ha, J.-W., Eom, J.-H., Kim, S.-C., and Zhang, B.-T. (2007). Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 2709–2716.

Ha, J.-W., Kim, S.-J., and Zhang, B.-T. (2013). Bayesian evolutionary methods for learning higher-order graphical models from high-dimensional data. *IEEE Transactions on Evolutionary Computation*. (in revision).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Han, B., Li, L., Chen, Y., Zhu, L., and Dai, Q. (2011). A two step method to identify clinical outcome relevant genes with microarray data. *J Biomed Inform*, 44(2):229–238.

Han, L., Witmer, P., Casey, E., Valle, D., and Sukumar, S. (2007). DNA methylation regulates microRNA expression. *Cancer Biology & Therapy*, 6(8):1290–1294.

Handel, A., Ebers, G., and Ramagopalan, S. (2010). Epigenetics: molecular mechanisms and implications for disease. *Trends in Molecular Medicine*, 16(1):7–16.

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402:C47–C52.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.

Higgins, M., Claremont, M., Major, J., Sander, C., and Lash, A. (2007). CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Research*, 35(suppl 1):D721–D726.

Holland, J. H. (1980). Adaptive algorithms for discovering and using general patterns in growing knowledge bases. *International Journal of Policy Analysis and Information Systems*, 4(3):245–268.

Holschneider, C. and Berek, J. (2000). Ovarian cancer: epidemiology, biology, and prognostic factors. In *Seminars in Surgical Oncology*, volume 19, pages 3–10.

Hornberg, J., Bruggeman, F., Westerhoff, H., and Lankelma, J. (2006). Cancer: a systems biology disease. *Biosystems*, 83:81–90.

Hu, H., Yan, X., Huang, Y., Han, J., and Zhou, X. (2005). Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21(suppl 1):i213–i221.

Hu, T., Xiong, H., Zhou, W., Sung, S. Y., and Luo, H. (2008). Hypergraph partitioning for document clustering: A unified clique perspective. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 871–872.

Huang, E., Cheng, S. H., Dressman, H., et al. (2003). Gene expression predictors of breast cancer outcomes. *Lancet*, 361:1590–1596.

Huang, J., Morris, Q., and Frey, B. (2006). Detecting microRNA targets by linking sequence, microRNA and gene expression data. In *Research in Computational Molecular Biology*, pages 114–129.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of Bioinformatics and Computational Biology*, 2(01):77–98.

Ivan, A., Halfon, M., Sinha, S., et al. (2008). Computational discovery of cis-regulatory modules in drosophila without prior knowledge of motifs. *Genome Biology*, 9(1):R22.

Jemal, A., Siegel, R., and Ward, E. (2010). Cancer statistics, 2010. *CA: A Cancer Journal for Clinicians*, 60:277–300.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Research*, 37(suppl 1):D98–D104.

Jones, P. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

Jones, P. and Baylin, S. (2007). The epigenomics of cancer. *Cell*, 128(4):683–692.

Joung, J.-G. and Fei, Z. (2009). Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model. *Bioinformatics*, 25(3):387–393.

Joung, J.-G., Hwang, K.-B., Nam, J.-W., Kim, S.-J., and Zhang, B.-T. (2007). Discovery of microRNA–mRNA modules via population-based probabilistic learning. *Bioinformatics*, 23(9):1141–1147.

Joung, J.-G., Kim, D., Kim, K., and Kim, J. (2013). Extracting coordinated patterns of DNA methylation and gene expression in ovarian cancer. *Journal of the American Medical Informatics Association*, 20(4):637–642.

Joung, J.-G., Kim, S.-J., Shin, S.-Y., and Zhang, B.-T. (2012). A probabilistic coevolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset. *BMC Bioinformatics*, 13(Suppl 17):S12.

Khatri, P., Voichita, C., Kattan, K., Ansari, N., Khatri, A., Georgescu, C., Tarca, A., and Draghici, S. (2007). Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Research*, 35(suppl 2):W206–W211.

Kim, D., Shin, H., Song, Y. S., and Kim, J. H. (2012a). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform*, 45(6):1191–1198.

Kim, S., Kim, S.-J., and Zhang, B.-T. (2007). Evolving hypernetwork classifiers for microrna expression profile analysis. In *Proceedings of IEEE Congress on Evolutionary Computation (CEC)*, pages 313–319.

Kim, S.-J., Ha, J.-W., Lee, B., and Zhang, B.-T. (2010). Evolutionary layered hypernetworks for identifying microrna-mrna regulatory modules. In *Proceedings of IEEE World Congress Computational Intelligence (WCCI-CEC)*, pages 2299–2306.

Kim, S.-J., Ha, J.-W., and Zhang, B.-T. (2012b). Identifying functional miRNA-mRNA modules based on hypergraph-based learning. In *Proceedings of IEEE International Student Paper Contest, Seoul Section*, pages 73–78.

Kim, S.-J., Ha, J.-W., and Zhang, B.-T. (2013a). Bayesian evolutionary hypergraph learning for predicting cancer clinical outcomes. *Journal of Biomedical Informatics*. (in revision).

Kim, S.-J., Ha, J.-W., and Zhang, B.-T. (2013b). Constructing higher-order miRNA-mRNA interaction networks in prostate cancer via hypergraph-based learning. *BMC Systems Biology*, 7(1):47.

Kim, S.-J., Joung, J.-G., and Zhang, B.-T. (2006). Co-evolutionary biclustering for microrna expression profiles analysis. In *Proceedings of the 7th International Conference of Korean Society for Bioinformatics*, pages 60–65.

Kindermann, R. and Snell, J. (1980). *Markov random fields and their applications*, volume 1. American Mathematical Society Providence, RI.

Klamt, S., Haus, U., and Theis, F. (2009). Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385.

Kojima, S., Chiyomaru, T., Kawakami, K., et al. (2011). Tumour suppressors miR-1 and miR-133a target the oncogenic function of purine nucleoside phosphorylase (PNP) in prostate cancer. *British Journal of Cancer*, 106(2):405–413.

Kok, S. and Domingos, P. (2009). Learning markov logic network structure via hypergraph lifting. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 505–512.

Kollar, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. The MIT Press.

Koziol, J. A., Feng, A. C., Jia, Z., Wang, Y., Goodison, S., McClelland, M., and D.Mercola (2009). Ensemble tree classifiers for prostate cancer prognosis. *Bioinformatics*, 25(1):54–60.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

Krogh, A., Larsson, B., Heijne, G., and Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580.

Kypta, R. and Waxman, J. (2012). Wnt/β-catenin signalling in prostate cancer. *Nature Reviews Urology*, 9(8):418–428.

Ladetto, M., Vallet, S., Trojan, A., et al. (2005). Cyclooxygenase-2 COX-2 is frequently expressed in multiple myeloma and is an independent predictor of poor outcome. *Blood*, 105(12):4784–4791.

Laird, P. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J., Armananzas, R., Santafe, G., Perez, A., and Robles, V. (2006). Machine learning in bioinformatics. *Brief Bioinform*, 7(1):86–112.

Lee, W. and Tzou, W. (2009). Computational methods for discovering gene networks from expression data. *Brief Bioinform*, 10(4):408–423.

Lehar, J., Krueger, A., Zimmermann, G., and Borisy, A. (2008). High-order combination effects and biological robustness. *Molecular Systems Biology*, 4(215):1–6.

Leva, G. D. and Croce, C. (2013). The role of microRNAs in the tumorigenesis of ovarian cancer. *Frontiers in Oncology*, 3.

Li, H. and Zhang, R. (2013). Role of EZH2 in epithelial ovarian cancer: from biological insights to therapeutic target. *Frontiers in Oncology*, 3.

Li, W., Zhang, S., Liu, C., and Zhou, X. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–2466.

Li, X., Gill, R., Cooper, N., Yoo, J., and Datta, S. (2011). Modeling microrna-mrna interactions using pls regression in human colon cancer. *BMC Med Genomics*, 19(4):44.

Li, Z. (1995). *Markov random field modeling in computer vision*. Springer-Verlag New York, Inc.

Li, Z. (2009). *Markov random field modeling in image analysis*. Springer.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740.

Liu, B., Li, J., Tsykin, A., Liu, L., Gaur, A., and Goodall, G. (2009a). Exploring complex miRNA-mRNA interactions with bayesian networks by splitting-averaging strategy. *BMC Bioinformatics*, 10(1):408.

Liu, B., Li, J., and Tsykin, T. (2009b). Discovery of functional miRNA-mRNA regulatory modules with computational methods. *Journal of Biomedical Informatics*, 42(4):685–691.

Liu, B., Liu, L., Tsykin, A., Goodall, G., Green, J., Zhu, M., Kim, C., and Li, J. (2010). Identifying functional miRNA–mrna regulatory modules with correspondence latent Dirichlet allocation. *Bioinformatics*, 26(24):3105–3111.

Liu, Y., Qiao, N., Zhu, S., Su, M., Sun, N., Boyd-Kirkup, J., and Han, J. (2013). A novel Bayesian network inference algorithm for integrative analysis of heterogeneous deep sequencing data. *Cell research*.

Liu, Z., Wang, Y., Zhang, X., and Chen, L. (2012). Network-based analysis of complex diseases. *IET Systems Biology*, 6(1):22–33.

Lo, Y., Yu, J., Chen, S., Hsu, G., Mau, Y., Yang, S., Wu, P., and Shen, C. (2007). Breast cancer risk associated with genotypic polymorphism of the mitotic checkpoint genes: a multigenic study on cancer susceptibility. *Carcinogenesis*, 28(5):1079–1086.

Lu, L., Katsaros, D., IA, I. L., Sochirca, O., and Yu, H. (2007). Hypermethylation of let-7a-3 in epithelial ovarian cancer is associated with low insulin-like growth factor-II expression and favorable prognosis. *Cancer research*, 67(21):10117–10122.

Lu, Y., Zhou, Y., Qu, W., Deng, M., and Zhang, C. (2010). A lasso regression model for the construction of microrna-target regulatory networks. *Bioinformatics*, 27(17):2406–2413.

Lujambio, A., Calin, G., Villanueva, A., et al. (2008). A microRNA DNA methylation signature for human cancer metastasis. *Proceedings of the National Academy of Sciences*, 105(36):13556–13561.

Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Marx, A., Kahan, T., and Simon, I. (2013). Integrative analysis of methylome and transcriptome reveals the importance of unmethylated CpGs in non CpG island gene activation. *Biomed Res Int*, 2013:785731.

Mason, O. and Verwoerd, M. (2007). Graph theory and networks in biology. *Systems Biology, IET*, 1(2):89–119.

Matsui, S., Ito, M., Nishiyama, B., Uno, H., Kotani, H., Watanabe, J., Guilford, P., Reeve, A., Fukushima, M., and Ogawa, O. (2007). Genomic characterization of multiple clinical phenotypes of cancer using multivariate linear regression models. *Bioinformatics*, 23(6):732–738.

McGovern, S., Qi, Y., Pusztai, L., et al. (2012). Centromere protein-A, an essential centromere protein, is a prognostic marker for relapse in estrogen receptor-positive breast cancer. *Breast Cancer Res*, 14:R72.

McKinney, B., Reif, D., Ritchie, M., and Moore, J. (2006). Machine learning for detecting gene-gene interactions: a review. *Appl Bioinformatics*, 5(2):77–88.

Mitra, K., Carvunis, A., Ramesh, S., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

Moore, J. and Ritchie, M. (2004). The challenges of whole-genome approaches to common diseases. *Journal of the American Medical Association*, 291:1642–1643.

Murphy, K. (2012). *Machine learning: a probabilistic perspective*. The MIT Press.

Nasser, S., Cunliffe, H., Black, M., and Kim, S. (2010). Context-specific gene regulatory networks subdivide intrinsic subtypes of breast cancer. *BMC Bioinformatics*, 29(12):S3.

Neapolitan, R. (2004). *Learning Bayesian networks*. Pearson Prentice Hall Upper Saddle River.

Nikovski, D. (2000). Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 12(4):509–516.

Nunez-Iglesias, J., Liu, C., Morgan, T., Finch, C., and Zhou, X. (2010). Joint genome-wide profiling of miRNA and mRNA expression in alzheimer's disease cortex reveals altered miRNA regulation. *PLoS One*, 5(2):e8898.

Pearce, C., Near, A., Butler, J., et al. (2008). Comprehensive evaluation of ESR2 variation and ovarian cancer risk. *Cancer Epidemiology Biomarkers & Prevention*, 17(2):393–396.

Peng, X., Guo, W., Liu, T., et al. (2011). Identification of miRs-143 and-145 that is associated with bone metastasis of prostate cancer and involved in the regulation of EMT. *PLoS One*, 6(5):e20341.

Peng, X., Li, Y., Walters, K., Rosenzweig, E., Lederer, S., Aicher, L., Proll, S., and Katze, M. (2009). Computational identification of hepatitis c virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, 10(1):373.

Philips, S., Richter, A., Oesterreich, S., et al. (2012). Functional characterization of a genetic polymorphism in the promoter of the ESR2 gene. *Hormones and Cancer*, 3(1-2):37–43.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Ramadan, E., Perincheri, S., and Tuck, D. (2010). A hypergraph approach for analyzing transcriptional networks in breast cancer. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 556–562.

Raychaudhuri, P. and Park, H. (2011). FoxM1: a master regulator of tumor metastasis. *Cancer Research*, 71(13):4329–4333.

Roddick, J. F., Spiliopoulou, M., Lister, D., and Ceglar, A. (2008). Higher-order mining. *ACM SIGKDD Explorations Newsletter*, 10(1):5–17.

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517.

Schaefer, A., Jung, M., Mollenkopf, H., Wagner, I., Stephan, C., Jentzmik, F., Miller, K., Lein, M., Kristiansen, G., and Jung, K. (2010). Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. *International Journal of Cancer*, 126(5):1166–1176.

Scharer, C., McCabe, C., Ali-Seyed, M., Berger, M., Bulyk, M., and Moreno, C. (2009). Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer Research*, 69(2):709–717.

Schlitt, T. and Brazma, A. (2007). Current approaches to gene regulatory network modeling. *BMC Bioinformatics*, 8(6):S9.

Schwartz, S., Caceres, C., Morote, J., Torres, I. D., Rodriguez-Vallejo, J., Gonzalez, J., and Reventos, J. (1999). Gains of the relative genomic content of ErbB1 and ErbB2 in prostate carcinoma and their association with metastasis. *Int J Oncol*, 14(2):367–371.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, 36:1090–1098.

Segal, E., Peer, D., Regev, A., Koller, D., and Friedman, N. (2005). Learning module networks. *Journal of Machine Learning Research*, 6:557–588.

Segal, E., Shapira, M., Regev, A., Peer, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176.

Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, 17(suppl 1):S243–S252.

Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y. (2007). Global and local architecture of the mammalian microrna transcription factor regulatory network. *PLoS Comput Biol*, 3(7):e131.

Shi, L., Campbell, G., Jones, W. D., et al. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827–838.

Siegfried, Z. and Simon, I. (2010). DNA methylation and gene expression. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(3):362–371.

Simon, R. (2003). Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*, 89:1599–1604.

Smoot, M., Ono, K., Ruscheinski, J., Wang, P., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Spisák, S., Kalmár, A., Galamb, O., et al. (2012). Genome-wide screening of genes regulated by DNA methylation in colon cancer development. *PLoS One*, 7(10):e46215.

Steele, R., Mott, J., and Ray, R. (2010). MBP-1 upregulates miR-29b, which represses Mcl-1, collagens, and matrix metalloproteinase-2 in prostate cancer cells. *Genes & Cancer*, 1(4):381–387.

Street, N., Mangasarian, O. L., and Wolberg, W. H. (1995). An inductive learning approach to prognostic prediction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 522–530.

Sun, B. Y., Zhu, Z. H., Li, J., and Linghu, B. (2011). Combined feature selection and cancer prognosis using support vector machine regression. *IEEE ACM Trans Comput Biol Bioinform*, 8(6):1671–1677.

Taneja, P., Maglic, D., Kai, F., Zhu, S., Kendig, R., Fry, E., and Inoue, K. (2010). Classical and novel prognostic markers for breast cancer and their clinical significance. *Clinical Medicine Insights Oncology*, 4:15.

Taylor, B., Schultz, N., Hieronymus, H., et al. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell*, 18:11–22.

Tian, Z., Hwang, T., and Kuang, R. (2008). A hypergraph-based learning algorithm for classifying gene expression and arraycgh data with prior knowledge. *Bioinformatics*, 25(21):2831–2838.

Tran, D., Satou, K., and Ho, T. (2008). Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics*, 9(Suppl 12):S5.

Triulzi, T., Iorio, M., Tagliabue, E., and Casalini, P. (2013). microRNA: New players in metastatic process. *Oncogene and Cancer-From Bench to Clinic*, page 391.

Upstill-Goddard, R., Eccles, D., Fliege, J., and Collins, A. (2013). Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief Bioinform*, 14(2):251–260.

van Eijk, K., de Jong, S., Boks, M., et al. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, 13(1):636.

Veer, L. J. v., Dai, H., van de Vijver, M. J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.

Verduijn, M., Peek, N., Rosseel, P. M., de Jonge, E., and de Mol, B. A. (2007). Prognostic bayesian networks i: rationale, learning procedure, and clinical use. *J Biomed Inform*, 40(6):609–618.

Volinia, S., Galasso, M., Costinean, S., et al. (2010). Reprogramming of mirna networks in cancer and leukemia. *Genome Res*, 20:589–599.

Von, L. U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Vrba, L., Jensen, T., Garbe, J., Heimark, R., Cress, A., Dickinson, S., Stampfer, M., and Futscher, B. (2010). Role for DNA methylation in the regulation of miR-200c and miR-141 expression in normal and cancer cells. *PLoS One*, 5(1):e8697.

Wang, C., Komodakis, N., and Paragios, N. (2013). Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627.

Wang, E., Lenferink, A., and O'Connor-McCourt, M. (2007). Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol Life Sci*, 64(14):1752–1762.

Watahiki, A., Wang, Y., Morris, J., Dennis, K., O'Dwyer, H., Gleave, M., Gout, P., and Wang, Y. (2011). MicroRNAs associated with metastatic prostate cancer. *PLoS One*, 6(9):e24950.

Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544.

Wheeler, T., Clements, J., Eddy, S., et al. (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*, 41(D1):D70–D82.

Wolfer, A. and Ramaswamy, S. (2011). MYC and metastasis. *Cancer Research*, 71(6):2034–2037.

Wong, K., Huang, X., and Chim, C. (2012). DNA methylation of microRNA genes in multiple myeloma. *Carcinogenesis*, 33(9):1629–1638.

Wu, Q., Guo, R., Lin, M., Zhou, B., and Wang, Y. (2011). MicroRNA-200a inhibits CD133/1+ ovarian cancer stem cells migration and invasion by targeting E-cadherin repressor ZEB2. *Gynecologic oncology*, 122(1):149–154.

Yan, H., Choi, A.-J., Lee, B., and Ting, A. (2011). Identification and functional analysis of epigenetically silenced microRNAs in colorectal cancer cells. *PLoS One*, 6(6):e20628.

Yan, X., Mehan, M., Huang, Y., Waterman, M., Yu, P., and Zhou, X. (2007). A graph-based approach to systematically reconstruct human transcriptional regulatory modules. *Bioinformatics*, 23(13):i577–i586.

Yoon, S. and Micheli, G. D. (2005). Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, 21(suppl 2):ii93–ii100.

Zhang, B.-T. (2008). Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. *IEEE Computational Intelligence Magazine*, 3(3):49–63.

Zhang, H., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A. (2012a). AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 40(D1):D144–D149.

Zhang, S., Li, Q., Liu, J., and Zhou, X. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):401–409.

Zhang, S., Liu, C., Li, W., Shen, H., Laird, P., and Zhou, X. (2012b). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391.

Zhang, W., Edwards, A., Fan, W., Flemington, E., and Zhang, K. (2012c). miRNA-mRNA correlation-network modules in human prostate cancer and the differences between primary and metastatic tumor subtypes. *PLoS One*, 7(6):e40130.

Zhou, D., Huang, J., and Schoelkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608.

Zhu, J., Jiang, Z., Gao, F., et al. (2011). A systematic analysis on DNA methylation and the expression of both mRNA and microRNA in bladder cancer. *PLoS One*, 6(11):e28223.

초 록

생물학 시스템의 포괄적 이해는 많은 유전적 인자들간의 고차 상호작용 분석을 필요로 한다. 다양한 유전적 인자들은 협력하여 암 발생, 진행, 전이를 포함한 생물학적 프로세스에 영향을 준다. 그러나 유전적 상호작용의 복잡성 때문에 수많은 유전적 인자들간의 동시 조절 역할 및 기능적 영향 발굴에 대한 연구는 여전히 어려운 문제로 인식되고 있다. 본 논문에서는 암을 포함한 생물학적 프로세스에서의 수많은 유전적 인자들간 복잡한 관계를 분석하기 위한 하이퍼그래프 모델을 제안한다. 첫째, 고차 유전적 상호작용 표현이 명확하게 가능한 새로운 하이퍼그래프 기반 모델을 제안한다. 둘째, 고차 관계로 구성된 거대한 문제공간을 효율적으로 탐색 할 수 있는 진화연산 기반 하이퍼그래프 학습 방법을 제안한다. 마지막으로 학습된 하이퍼그래프 모델로부터 생물학적 지식을 효과적으로 추출 할 수 있는 방법을 제안한다.

하이퍼그래프 모델은 고차원 데이터로부터 수많은 변수간의 복잡한 관계를 명확하게 표현 가능한 고차 그래프 모델이다. 이러한 특성은 다양한 유전적 인자들간 고차 상호작용이 특징적으로 나타나는 생물학 및 의학 현상 분석에 적합하다. 이에 본 논문에서는 유전체 수준에서 동시 조절 유전적 상호작용 발굴에 초점을 맞추어 대규모 생물학 데이터 분석을 위한 학습 기법과 모델 구조 관점에서 향상된 하이퍼그래프 모델을 제시한다. 제안된 모델 학습 기법은 인자간 고차 관계를 표현하는 광범위한 복잡한 문제공간을 더욱 효율적으로 탐색하기 위해 기존의 진화 학습 과정에서 정보이론적 기준을 도입하여 학습을 위한 정책으로 활용하며, 이러한 진화 학습 과정은 순차적 베이지안 샘플링 프레임워크를 기반으로 설명된다. 또한, 복잡한 생물학 기전을 반영하여 계층적 유전적 관계를 모델링 하기 위해 하이퍼그래프 모델에 계층 구조를 도입한다. 이러한 계층적 구조는 후성유전적, 전사적, 전사후 과정에 걸친 기능적 블럭 또는 그룹으로의 유전자 규제 회로를 더욱 명확하게 표현 가능하게 한다. 또한 학습된 모델 구조를 분석 할 수 있는 그래프 분석 방법을 제시하여 유전적 모듈 및 조절 네트워크와 같은 생물학적 시스템들의 포괄적인 구조 파악이 가능하도록 한다.

제안한 모델은 현 생물학 및 의학에서 주요 주제로 간주되고 있는 암 유전체 분석에 응용한다. 실험 결과를 통해 다양한 암 유전체 데이터에서 제안한 모델 성능이 다른 최신 기술 수준 모델에 비등하거나 더 우수함을 확인하였다. 이에 더 나아가 제안한 모델은 특정 암과 연관된 유전자 모듈, miRNA-mRNA 네트워크, 복합 유전적 상호작용과 같은 새롭거나 숨겨진 패턴의 잠재적 유전자 회로의 후보군을 발견 할 수 있다. 이러한 분석의 결과는 암 기전에서 아직 알려지지 않은 기능을 발굴하는데 있어 핵심적인 증거들을 제공 할 수 있다. 이에 제안한 하이퍼그래프 모델은 암을 포함한 생물학적 프로세스의 핵심 조절 메커니즘을 밝히고, 더불어 그것의 포괄적인 이해에 공헌 할 수 있을 것이다.

Keywords: **하이퍼그래프, 고차 그래프 모델, 진화 학습, 유전적 상호작용, 유전자 모듈, miRNA-mRNA 네트워크, 암**

학번: 2005-20561