理學博士學位論文

# Evolutionary Machine Learning of Higher Order Relationships in Genome-wide Sequence Analysis

유전체 서열 분석에서 고차 관계의 진화적 기계학습

2014年 2月

서울大學校 大學院

협동과정 생물정보학 전공

李 齊 根

# 유전체 서열 분석에서 고차 관계의 진화적 기계학습

# (Evolutionary Machine Learning of Higher Order Relationships in Genome-wide Sequence Analysis)

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 전공

李 齊 根

李齊根의 理學博士 學位論文을 認准함

2013年 12月

委 員 長 김 선

副委員長 장 병 탁

委 員 김 희 발

委 員 신 수 용

委 員 백 대 현

# 유전체 서열 분석에서 고차 관계의 진화적 기계학습

# Evolutionary machine learning of higher order relationships in genome-wide sequence analysis

Je-Keun Rhee

Ph.D. Thesis

Interdisciplinary Program in Bioinformatics
Seoul National University
Feb., 2014

Supervisor: Byoung-Tak Zhang

# Abstract

One of the basic research goals in life science is to understand the complex relationships between biological factors and phenotypes, and to identify the various factors affecting the phenotype. In particular, genomic sequences play a significant role in determining the phenotype, such as gene expression and a susceptibility to disease, so the studies for the fundamental information stored in genome is essential to understanding biological processes. Previous genomic sequence analyses mainly focused on identification of a single associated factor or pairwise relationships with significant effects. Recent development of high-throughput technologies has made it possible to identify the causal factors by carrying out genome-wide analysis. However, it still remains as a challenge to discover higher-order interactions of multiple factors because this involves huge search spaces and computational costs.

In this dissertation, we develop effective methods for identifying the higher-order relationships of sequence elements affecting the phenotype, by combining statistical learning with evolutionary computation. The methods are applied to finding the associated combinatorial factors and dysfunctional modules in various genome-wide sequence analysis problems. Firstly, we show statistical learning-based methods to detect co-regulatory sequence motifs and to investigate combinatorial effects of DNA methylation, affecting on downstream gene expression. Next, to examine the sequence datasets with a huge number of attributes on human genome, we apply evolutionary computation approaches. Our methods search the problem feature space based on machine learning techniques using training datasets in evolutionary computation processes and are able to find candidate solution well in computationally expensive optimization problems. The experimental results show that the approaches are useful to find the higher-order relationships associated to disease using genomic and epigenomic datasets. In conclusion, our studies would provide practical methods to analyze complex interactions among sequence elements in genomic/epigenomic studies.

**Keywords: Higher-order interaction, Evolutionary computation, Genome-wide sequence analysis, Machine learning, Genomics, Epigenomics**

**Student Number: 2004-20623**

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

The post-genomic era is characterized by a tremendous revolutionary expansion in biological data. Over past few decades, there has been rapid development in biological research and technologies, as a result, a huge amount of data have been produced. In particular, with the advances of sequencing technologies, a large amount of datasets have been deposited in repositories (Metzker, 2009). Understanding and exploiting these data is now a key to success of advancing biological research, and the requirements have stimulated development and expansion of applying computational approaches in biology.

The large expansion of genome-wide measurement data poses the research question of how to retrieve the valuable knowledge from the genomic sequences (Huttenhower and Hofmann, 2010; Chin et al., 2011). Traditionally, genomic studies mainly focused on central dogma in molecular biology, namely from genome to transcriptome. Experimentally determined catalogues of genes only tell us about a basic building block of the biological regulatory processes. They do not tell us much about how the biological processes operate as a system, such as higher order functional behaviors (Chuang et al., 2010). Although many computational approaches have dealt with high-throughput biological datasets generated in multi-dimensional forms, it is still important to search the large datasets efficiently and effectively (Palsson and Zengler, 2010; Kouskoumvekaki et al., 2013).

Actually, most biological problems are complex and hard to be understood. One problem is to investigate the interactions of the various factors, since the biological processes are affected by multiple factors. Although genome-wide analysis can be possible with the development of high-throughput technologies, an exhaustive search of all potential solutions is still challenging, and most likely impossible. The standard constructive and approximate approaches are usually impractical in terms of a huge search space and lots of computational costs. Thus, the genome-wide sequence analyses mainly focused on identification of a single associated factor or pairwise relationships with significant main effects (Cordell, 2009; Bush and Moore, 2012).

The genome-wide sequence analyses have contributed to ability to identify genomic sequence elements that are associated with phenotypes such as gene expression and disease (Feero et al., 2010; Heap et al., 2009; Kim et al., 2012; Kang et al., 2011). It has been possible to find a single sequence element that has statistically significant association with phenotype. To date, many associated genes or sequence elements were found, but there were not clear explain the complex biological processes (McCarthy et al., 2008; Stranger et al., 2011). Thus, there has been increased interest in discovering combinations of the sequence elements that are strongly associated with a phenotype even if each element has little or even no individual effect. Thus, an alternative research view of post-genomic/epigenomic era would be to go up eventually to still higher levels, i. e. biological systems.

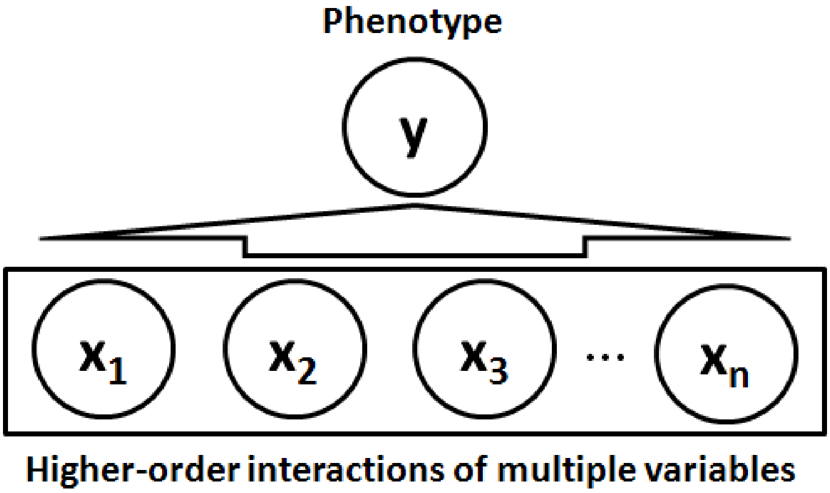


Figure 1.1: Schematic concept for higher-order interaction and its effects on phenotype

## 1.2 Approaches

In this dissertation, we present computational approaches to predict higher-order relationships of disease genes or sequence elements and identify dysfunctional modules, based on machine learning and evolutionary computation using phenotype and sequence information. Our goal is to discover and study the combinations of sequence elements affecting on phenotype. In particular, we focus on discovering the interactions, especially high-order ones beyond size 2, that are strongly associated with a phenotype and yield information on interpretable statistical and functional interactions (Figure 1.1).

At first, we show a way to search co-regulatory sequence motifs using a statistical learning method, kernel canonical correlation analysis (kernel CCA) (Rhee et al., 2009). One of the major challenges in gene regulation studies is to identify regulators affecting the expression of their target genes in specific biological processes. Despite their importance, regulators involved in diverse biological processes still remain largely unrevealed. In the study, we propose a kernel-based approach to efficiently identify core regulatory elements and their combinations involved in specific biological processes using gene expression profiles. We develop a framework that can detect correlations between gene expression profiles and the upstream sequences on the basis of the kernel canonical correlation analysis (kernel CCA). We show that upstream sequence patterns are closely related to gene expression profiles based on the canonical correlation scores obtained by measuring the correlation between them. The experimental results show that our method is able to successfully identify regulatory motifs and their co-regulatory pairs involved in specific biological processes.

Secondly, we investigated the combinatorial effects of DNA methylation on downstream gene expression using machine learning approaches (Rhee et al., 2013). Aberrant DNA methylation of CpG islands (CGIs), CGI shores, and first exons is known to play a key role in the altered gene expression patterns in all human cancers. To date, a systematic study on the effect of DNA methylation on gene expression using high resolution data has not been reported. In this study, we conducted an integrated analysis of MethylCap-sequencing data and Affymetrix gene expression microarray data for 30 breast cancer cell lines representing different breast tumor phenotypes. We develop methylome data analysis protocols for the integrated analysis of DNA methylation and gene expression data on the genome scale and we present comprehensive genome-wide methylome analysis results for differentially methylated regions and their potential effect on gene expression in 30 breast cancer cell lines representing three molecular phenotypes, luminal, basal A, and basal B. Our integrated analysis demonstrates that methylation status of different genomic regions may play a key role in establishing transcriptional patterns in molecular subtypes of human breast cancer.

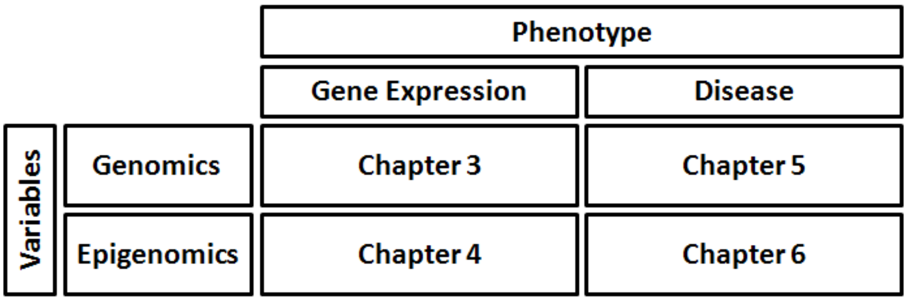
These two genome-wide approaches were useful for identification of co-regulatory interactions or combinatorial effects, associated to downstream gene expression. However, sometimes, it might need another approach to examine a huge number of sites on whole genome and to discover higher-order relationships of sequence elements associated with complex disease. Then, we applied evolutionary computation approaches to identify higher-order interaction of multiple factors associated to disease. Evolutionary computation is a general purpose search approach that uses principles inspired by natural genetic populations to evolve solutions to problems (Simon, 2013). The basic idea is to maintain a population of individuals which represent plausible solutions to the problem, which evolves over time through a process of competition and controlled variations.

In the framework of evolutionary machine learning, the main idea is that the evolutionary computation method has stored training data to search problem feature space and population information during the iterative evolutionary process. Then, the machine learning technique is helpful in analyzing these data for enhancing the search performance.

We propose an approach to search the higher-order interaction for genome-wide association studies based on the evolutionary machine learning. Searching for the relationship between the genetic variant and its phenotypic effects is important to understand the genetic basis and mechanism of many complex genetic diseases. There have been a lot of research to analyze the causality and, in many studies, it have led to succeed to discover the associations of genes with diseases. Although there exist lots of the genetic variants with major effects and they can be linked to complex diseases, however, it is still challenging to find the multiple interactions from a millions of SNPs and their association with a disease. Here, we present an approach to analyze higher-order interactions of the genetic variations, which associated with a disease. The method searches combinatorial feature spaces of the genetic variants and selects the higher-order variables which are distinctive to classify the disease and normal samples by evolutionary learning. We test the method and illustrate the advantages with genetic variant datasets for type 2 diabetes. As a result, our approach could identify the higher-order interaction of SNPs associated with type 2 diabetes, and especially detect several interactions specific in Korean population.

Finally we introduced probabilistic concepts in the evolutionary computation for identification of DNA methylation modules. By exploring the problem space by building and sampling explicit from probabilistic graphical models, the approach would be proper to find the higher-order relationships or biological modules.

Considerable studies have been made to elucidate effects of genetic variability in complex disease, but it is still challenging to discover molecular pathogenesis clearly. The epigenetic factor would be another candidate to make up the complex regulatory mechanism. Especially it is well-known that DNA methylation could lead to inhibition of downstream gene expression. Although many researchers are trying to clarify the relationships between DNA methylation and gene expression, recently, more efforts are required to find the multiple interactions from a lot of DNA methylation sites and their association with a disease. To assess DNA methylation modules potentially relevant to disease, we use an estimation of distribution algorithm (EDA)-based learning method identifying high-order interaction of DNA methylation sites. It finds a solution which is a set of discriminative methylation sites by building a probabilistic dependency model. The algorithm is applied to array- and sequencing-based high-throughput DNA methylation profiling datasets, and the experimental



| | | Phenotype | |
|---|---|---|---|
| | | Gene Expression | Disease |
| Variables | Genomics | Chapter 3 | Chapter 5 |
| | Epigenomics | Chapter 4 | Chapter 6 |

Figure 1.2: Organization of chapters

results show that it has a good search ability to identify the DNA methylation modules for a specific disease.

Our approaches would provide practical methods to integrate large amount of datasets and to analyze complex interactions among building blocks and with dynamic environments.

## 1.3 Organization of the dissertation

This dissertation is organized as follows (Figure 1.2):

- In Chapter 2, we briefly introduces informatics and computational approaches in genomic analysis. We describe background of genome biology, and explain what the machine learning and evolutionary computation are. Then, the basic concepts and their several applications in biological domains are described.

- In Chapter 3, we search co-regulatory sequence motifs by a kernel-based correlation analysis. We identify regulatory sequences affecting the expression of their downstream genes. And we investigate pairwise relationships of the sequence motifs closely related to gene expression profiles in a specific biological process.

- Chapter 4 discribes analysis protocols to investigate effects of DNA methylation in various sites on downstream gene expression. Using high resolution sequencing-based methylation profiling datasets, we show comprehensive genome-wide methylome analysis results for their potential effect on gene expression. The analysis results present that methylation status of different genomic regions may play combinatorial effects on transcriptional patterns via a statistical learning approach.

- In Chapter 5, we propose an evolutionary learning method for identifying higher-order interaction of multiple SNPs in genome-wide association studies. We show that the proposed evolutionary learning method searches combinatorial feature spaces and identifies the higher-order variables which are related to disease.

- In Chapter 6, we use a probabilistic evolutionary learning to find higher-order relationships from a lot of DNA methylation sites, which is potentially relevant to disease. Instead of crossover or mutation operators in traditional evolutionary computation, we build a probabilistic distribution model and are sampled from the model in the evolutionary learning processes. The experimental method and results represent that the approach can be a new systematic way to identifying high-order interaction of DNA methylation sites and DNA methylation modules which is associated to disease.

- Finally, we summarize the dissertation and discuss our research in Chapter 7.

# Chapter 2

# Genome biology and computational analysis

## 2.1 Fundamentals of genome biology

### 2.1.1 DNA, gene, chromosomes and cell biology

DNA (deoxyribonucleic acid) is a biomolecule that includes information for how organisms are genetically built. DNA is a double strand structure that contains complementary genetic information encoded by 4 bases, adenine (A), guanine (G), thymine (T) and cytosine (C). A gene is a segment of DNA that can be inherited from parents to children and can confer a trait to the offspring. The genes are organized and packaged in chromosomes. In case of human, there exist 23 pairs of chromosomes.

One set of chromosomes for each pair comes from a person's mother, and the other set is from father. New cells get their chromosomes from old cells through cell division, mitosis. The chromosome in cell nucleus is divided into two identical sets by mitosis of cell cycle. The primary result of mitosis is the transferring of the parent cell's genome into two daughter cells. Cell cycle is the series of events leading to its growth, replication (duplication) and division of a eukaryotic cell. The cell cycle can be divided into several phases: G1, S, G2 and M phases. At G1 and G2 phases, cells increase in size and DNA replication occurs at S phase. M phase is a periods of mitosis which is cell division state. The cell growth stops at this stage and the cell divides itself into two distinct daughter cells.

### 2.1.2 Gene expression and regulation

Gene expression is a fundamental step at which a genotype gives rise to a phenotype. The gene expression means a process that the genetic information from a gene is used in production of a functional gene product (protein or RNA). The process is generally described by that a gene is transcribed into RNA and this transcript may then be translated into protein.

Regulation of gene expression includes mechanisms to increase or decrease the production of specific gene products. The program of gene expression is very sophisticate. A complex set of interactions between genes, RNA molecules, proteins (including transcription factors) and other components of the expression system determine when and where specific genes are activated and the amount of protein or RNA product produced. Some genes are expressed continuously, as they produce proteins involved in basic metabolic functions; some genes are expressed as part of the process of cell differentiation; and some genes are expressed as a result of cell differentiation.

Specific DNA sequences are accessible for specific proteins to bind. Many of these proteins are activators, while others are repressors. Such proteins are often called transcription factors (TFs). Transcription factors are proteins that play a role in regulating the transcription of genes by binding to specific regulatory nucleotide sequences. Each TF has a specific DNA binding domain that recognizes a 6-10 base-pair motif in the DNA, as well as an effector domain (Matys et al., 2003; Sandelin et al., 2004).

For an activating TF, the effector domain recruits RNA polymerase II, the eukaryotic mRNA-producing polymerase, to begin transcription of the corresponding gene. TFs bind at the promoters just upstream of eukaryotic genes. However, they also bind at regions called enhancers, which can be oriented forward or backwards and located upstream or downstream or even in the introns of a gene, and still activate or repress the gene expression. Studying gene expression across the whole genome via microarrays or massively parallel sequencing allows investigators to see which groups of genes are co-regulated during differentiation, cancer, and other states and processes.

### 2.1.3 Genomics

Genome is the entirety of all genes and information contained within the noncoding regions from an organism, mainly encoded by DNA. Genomics usually describe studies to determine the entire DNA sequence of organisms and genomic structures. The field also includes studies of various genomic phenomena. In contrast to the classical molecular biology or genetics to investigate the roles and functions of single gene, genomics aim to elucidate its effects on the entire genomic networks with its genetic and functional information (Lander, 1996).

A major branch of genomics is concerned with sequencing the genomes of various organisms. A rough draft of the human genome was completed in 2001 (Venter et al., 2001; Lander et al., 2001). Since then, there have been much more studies for human genome. Also, the genomic information of many other species has been successfully achieved. The knowledge of full genomes has created the possibility for the field of functional genomics, mainly concerned with patterns of gene expression during various conditions. For the purpose, computational approaches would be the most important tools here.

### 2.1.4 Epigenomics

The classical biology states that DNA is transcribed to RNA, RNA is translated to protein, and it regulates various cellular processes and functions. In the traditional views, phenotypic alteration has been caused by aberrant sequence variants or an inherited genomic allele. However, in the recent view, cells with identical DNA sequences can have a variety of distinct functions and phenotypes, by epigenetic modification including DNA methylation and histone modification (Murrell et al., 2005; Holliday, 2006). That is, the epigenetic modifications affect gene expression without altering the DNA sequences and play an important role in numerous cellular processes such as in differentiation, development and tumorigenesis (Bernstein et al., 2007; Baylin and Jones, 2011).

One of the most characterized epigenetic modifications is DNA methylation. DNA methylation is a process by which a methyl group is added to DNA. The methylation is most commonly found on cytosine residues adjacent to guanine, termed CpG dinucleotides (Laird, 2010). It is well-known that the DNA methylation can control gene expression. Usually the DNA methylation represses gene expression by a multistep process, although the exact mechanim is unknown.

Epigenomic research tries to identify and characterize epigenetic modifications on a global level. The study of epigenetics on a global level has been made possible recently through high-throughput assays. To manage a huge size of datasets and to clarify the complex mechanism on the fields, as in the other genomics fields, epigenomics also relies heavily on bioinformatics, which combines the disciplines of biology, mathematics and computer science.

## 2.2 Evolutionary machine learning

### 2.2.1 Machine learning and evolutionary computation

Machine learning is a study to give computers abilities to learn from existing data. Usually, it can be used to discover patterns and rules from data, and predict future events. Machine learning techniques generally involves statistical methods, interpolation and regression, supervised classification algorithms, clustering analysis, reinforcement learning, and so on.

The ideas and techniques from machine learning can be hybridized with evolutionary computation. Evolutionary computation with machine learning techniques would be a promising research direction to search optimal solution from the machine learning point of view (Zhang et al., 2011). Evolutionary computation is a kind of optimization methodology inspired by mechanisms of biological evolution. It can be widely used as an optimization tool in recent years.

The first step of the evolutionary computation is initialization of population. Next, it enters iterative evolutionary step with fitness evaluation, selection, and population reproduction. The newly generated population is evaluated again and the iteration continues until a termination criterion is satisfied.

### 2.2.2 Evolutionary computation in biology

The genomic revolution is generating a huge amount of data in rapid speed but it has become made difficult for biologists to decipher. In addition, many problems in biology are too large to solve with standard methods. Evolutionary computation can be a solution for the current bioinformatics problems (Fogel and Corne, 2002;

Pal et al., 2006). Although bioinformatics present a number of difficult optimization problems, evolutionary computation can rapidly search very large and complex spaces and return reasonable solutions.

The evolutionary computation has experienced a large growth in applications for bioinformatics with several advantages. For example, the errors generated in biological experiment data might be handled with no significant problem in the evolutionary computation. The errors can contribute to genetic diversity, a desirable property in the evolutionary learning processes. Thus, it might be more tolerable in using evolutionary computation than other deterministic algorithms. Sometimes, several tasks of bioinformatic studies do not require the exact optimum answer. Instead, they require robust and close approximate solutions. Also, local optimal solution can be helpful to understand biological processes. Evolutionary computation approaches can be also efficient to provide the solution in this case. In addition, EAs can process, in parallel, population billions times larger than is usual expectation is that larger populations can sustain larger range of genetic variation, and thus can generate high-fitness individuals in fewer generation. Laboratory operations on DNA inherently involve errors. These are more tolerable in executing evolutionary algorithms than executing deterministic algorithms.

Evolutionary computation has been profitably used in traditional bioinformatic problems. Several application areas follow:

- Sequence alignments

  Multiple sequence alignment helps to infer evolutionary history or discover conserved regions among closely related sequences. The problem is known as NP-hard. Genetic algorithms can be used to find optimal solutions in this problem (Notredame and Higgins, 1996; Nguyen et al., 2002).

- Motif finding

An instance of genetic algorithms can be used for motif finding, similar to Gibbs sampling. The motifs can generated from randomly selected sequences, and then alignment scores has been computed between the sequence fragments and the motifs. It increases the chance to find the real sequence motifs (Liu et al., 2004; Das and Dai, 2007).

- Protein structure prediction

  Evolutionary computation methods for protein structure prediction have been developed in the last decades. These have attempted to optimize the energy function of the peptide chain and to determine the optimal protein folding (Unger and Moult, 1993; Cooper et al., 2003).

- Protein-protein interaction and docking

  Protein interaction and docking represents fundamental function of biomolecules. Although it is possible now to determined by experimental methods, it is difficult to predict the recognition exactly ascertaining the structure of protein complexes. The evolutionary computation approaches can help to solve the problem (Morris et al., 1998; Wang et al., 2010).

The applications suggest that a variety of problems in biological domains can be well-suited for evolutionary computation approaches and be analyzed well by the methods.

# Chapter 3

# Identifying co-regulatory sequence motifs

## 3.1 Background

One of the major challenges in current biology is to elucidate the mechanism governing the gene expression. Gene expression programs depend mainly on transcription factors which bind to upstream sequences by recognizing short DNA motifs called transcription factor binding sites (TFBSs) to regulate their target gene expression (Lee et al., 2002). Transcription factors bind to upstream sequences to regulate gene expression. They recognize short DNA motifs called transcription factor binding sites (TFBSs). Although many regulatory motifs have been identified, large amount of functional elements still remain unknown (Xie et al., 2005).

Many genome-wide approaches have been developed in attempt to discover regulatory motifs from upstream sequences. The early computational approach for identifying regulatory motifs is based on statistical analyses using only upstream sequences of genes. Statistical methods such as maximum-likelihood estimation or Gibbs sampling, are effective for searching directly significant sequence motifs from multiple upstream sequences (Hughes et al., 2000; Bailey and Elkan, 1994). Several computational approaches based on machine learning methods have also been implemented. A SOM (self-organizing map)-based clustering method can find regulatory sequence motifs by grouping relevant sequence patterns (Mahony et al., 2005) and a graph-theoretic approach has tried to identify regulatory motifs by searching the maximum density subgraph (Fratkin et al., 2006).

More advanced approaches have been developed that can identify regulatory motifs by linking gene expression profiles and motif patterns. The main advantage of these approaches is that they can identify motifs correlated to specific biological processes. Most early trials used a unidirectional search, such as approaches that search for shared patterns with upstream sequences in a set of co-expressed genes that were found by clustering algorithms (Tavazoie et al., 1999; Brāzma et al., 1998) or those that determine whether genes with common regulatory elements are co-expressed (Pilpel et al., 2001; Park et al., 2002). In addition, it is also possible to link motifs to gene expression patterns using linear regression models or regression trees (Bussemaker et al., 2001; Keles et al., 2002). Recently, several techniques for a bidirectional search to detect the relationship between the regulatory motifs and the gene expression profiles have been emerged (Segal et al., 2003; Jeffery et al., 2007). They search regulatory motifs more efficiently than unidirectional approaches since they search similar expression patterns and regulatory motifs correlated to them simultaneously.

In this study, we propose a novel bidirectional approach using a kernel-based method, kernel CCA (kernel canonical correlation analysis), to analyze the relationship between regulatory sequences and gene expression profiles (Hardoon et al., 2004; Akaho, 2006; Bach and Jordan, 2003). The expression and sequence features are mapped from the original input space to a higher dimension space using a kernel trick, and the relationship between the two projected objects is interpreted to identify highly correlated motifs (Figure 3.1). Our method has advantages that it can detect core motifs relevant to a specific cellular process without the additional efforts of clustering and intensive motif sampling process in upstream sequences.

We applied the kernel CCA to a paired set of upstream sequence motifs of genes and their expression profiles in yeast *Saccharomyces cerevisiae* cell cycle, and explored significant relationships between motifs and expression profiles. We also searched for regulatory motifs correlated with specific expression patterns. We also searched for regulatory motifs correlated with specific expression patterns. Our method retrieved regulatory motifs that play an important role in cell cycle regulation including several well-known cell cycle regulatory motifs: MCB, SCB and SFF'. Furthermore, we identified motif pairs associated with the gene expression to construct a map of combinatorial regulation of regulators.

## 3.2 Methods

### 3.2.1 Investigation of the relationship between regulatory sequence motifs and expression profiles

Kernel CCA (Canonical correlation analysis) is a version of the nonlinear CCA, where the kernel trick is utilized to find nonlinearly correlated features from two datasets (Hardoon et al., 2004; Akaho, 2006; Bach and Jordan, 2003). Canonical correlation analysis (CCA) CCA is a classical multivariate statistical method for finding linearly correlated features from a pair of datasets (Hotelling, 1936). Suppose there is a pair of multivariates $\mathbf{x}$ and $\mathbf{y}$, CCA finds a pair of linear transformations such that the correlation coefficient between extracted features is maximized. How-

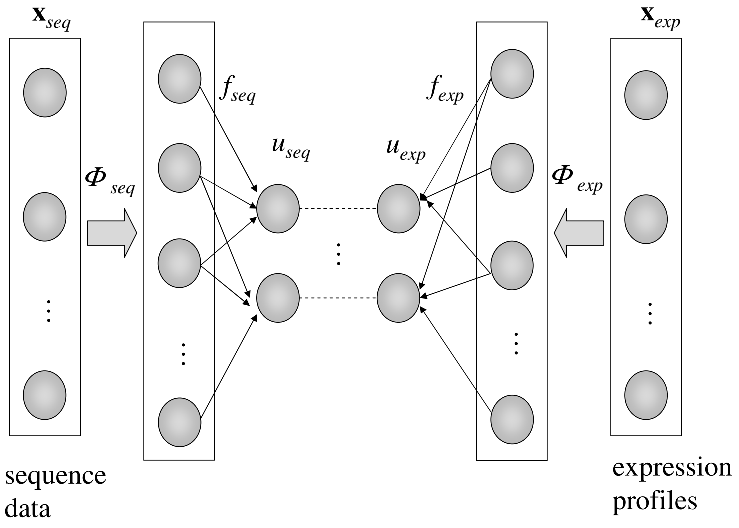


Figure 3.1: The basic scheme of the kernel CCA. The sequence and expression data are transformed to Hilbert space by $\phi$ function. By taking inner products, $u_{exp}$ and $u_{seq}$ were derived, which maximize the correlation between the upstream sequences and the expression profiles.

ever, if there is a nonlinear relationship between the variates, CCA does not always extract useful features.

Kernel CCA offers a solution for overcoming the linearity by first projecting the data into a higher dimensional feature space. While CCA is limited to linear features, kernel CCA can capture nonlinear relationships. Kernel CCA has been used for several applications including text retrieval and biological data analysis (Hardoon et al., 2004; Yamanishi et al., 2003).

Figure 3.1 illustrates the basic scheme of the kernel CCA for our integrated analysis of DNA sequence motif and gene expression data. Using kernel CCA, we tried to find maximally correlated features between the gene expression and the sequence motifs. Here, a gene set $\mathbf{X}$ is represented by two separate profiles in terms of its transcriptional behavior and upstream sequences, $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$. These are composed of the expression profile, $\mathbf{x}_{exp} = (e_1, e_2, ..., e_N)$ and the sequence profile, $\mathbf{x}_{seq} = (m_1, m_2, ..., m_M)$ of each gene. Here $e_i$ $(1 \leq i \leq N)$ is the expression value of the gene in the $i$-th sample or experimental condition from microarray data, and $m_j$ $(1 \leq j \leq M)$ denotes the occurrence frequency of the $j$-th sequence motif in the upstream region of the gene. For the detection of the correlated features between the two datasets, $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$ are first mapped to Hilbert space, $H$, by function $\phi$. That is, each $\mathbf{x}$ is projected into two directions, $f_{exp}$ and $f_{seq}$, in Hilbert space according to its representation:

$$u_{exp} = \Big\langle f_{exp}, \phi_{exp}(\mathbf{x}_{exp}) \Big\rangle \tag{3.1}$$

$$u_{seq} = \Big\langle f_{seq}, \phi_{seq}(\mathbf{x}_{seq}) \Big\rangle, \tag{3.2}$$

where $\Big\langle \cdot, \cdot \Big\rangle$ denotes the dot product. Kernel CCA looks for maximally correlated features between $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$:

$$\begin{aligned} &\gamma(f_{exp}, f_{seq}) = \\ &\quad \max \frac{\mathbf{cov}(u_{exp}, u_{seq})}{(\mathbf{var}(u_{exp}) + \lambda_{exp} \|f_{exp}\|^2)^{\frac{1}{2}} (\mathbf{var}(u_{seq}) + \lambda_{seq} \|f_{seq}\|^2)^{\frac{1}{2}}}, \end{aligned} \tag{3.3}$$

where $\lambda_{exp}$ and $\lambda_{seq}$ are regularization parameters. The kernel CCA can be given by solving a generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \mathbf{K}_{exp}\mathbf{K}_{seq} \\ \mathbf{K}_{seq}\mathbf{K}_{exp} & 0 \end{pmatrix} \begin{pmatrix} \alpha_{exp} \\ \alpha_{seq} \end{pmatrix} =$$
$$\rho \begin{pmatrix} (\mathbf{K}_{exp} + \frac{n\lambda_{exp}}{2}\mathbf{I})^2 & 0 \\ 0 & (\mathbf{K}_{seq} + \frac{n\lambda_{seq}}{2}\mathbf{I})^2 \end{pmatrix} \begin{pmatrix} \alpha_{exp} \\ \alpha_{seq} \end{pmatrix}, \tag{3.4}$$

where $\mathbf{I}$ denotes the identity matrix, $\mathbf{K}_{exp}$ is the kernel matrix for expression profile data, and $\mathbf{K}_{seq}$ is the kernel matrix for sequence motif data. When given $\alpha_{exp}$ and $\alpha_{seq}$ as the solution of the above generalized eigenvalue problem with the largest eigenvalue, canonical correlation scores (CC scores) for $\mathbf{x}_{seq}$ and $\mathbf{x}_{seq}$ are estimated by $u_{seq} = \mathbf{K}_{seq}\alpha_{seq}$ and $u_{exp} = \mathbf{K}_{exp}\alpha_{exp}$. The CC scores are the low dimensional mapping of genes in terms of two separate representations and can be used to show the salient correlation between the two. Once we have obtained the $\alpha$ vector, the weights of the motif and expression profile, $\mathbf{W}_{seq}$ and $\mathbf{W}_{exp}$, are obtained as follows:

$$\mathbf{W}_{exp} = \mathbf{x}_{exp}^T \alpha_{exp} \tag{3.5}$$

$$\mathbf{W}_{seq} = \mathbf{x}_{seq}^T \alpha_{seq}. \tag{3.6}$$

A high weight value of the specific sequence motif means that the motif is strongly correlated with the expression patterns of genes whose upstream region includes the motif and whose CC scores are high. If a weight of a specific motif has a high absolute value, the motif is more likely to be investigated further.

### 3.2.2 Preparation of the gene expression datasets

Expression profiles of all ORFs (open reading frames) during the yeast cell cycle that consists of 18 time points in the alpha factor synchronization case [18] were used as the expression dataset. To map from the expression profiles to high dimensional

Table 3.1: Known regulatory motifs in yeast *Saccharomyces cerevisiae.*

| Motif Name | | | |
|---|---|---|---|
| RAP1 | RPN4 | GCN4 | MCB |
| HAP234 | MIG1 | AFT1 | STRE' |
| CCA | CSRE | PHO4 | STE12 |
| HSE | ABF1 | ATRepeat | GAL |
| Leu3 | LYS14 | MET31-32 | OAF1 |
| PAC | PDR | PHO | REB1 |
| STRE | ECB | ndt80(MSE) | Yap1 |
| SCB | Gcr1 | zap1 | MCM1' |
| MCM1 | SFF | SFF' | BAS1 |
| Ume6(URS1) | SWI5 | ALPHA1' | ALPHA1 |
| ALPHA2' | ALPHA2 | | |

space, we converted them to the kernel matrix. We applied a gaussian RBF kernel to the expression profile matrix by: $k(\mathbf{x}_{exp}, \mathbf{x'}_{exp}) = \exp[-\frac{d(\mathbf{x}_{exp},\mathbf{x'}_{exp})}{2\sigma^2}]$, where $\sigma$ is a parameter and function $d(\cdot, \cdot)$ is a Euclidean distance.

### 3.2.3 Preparation of the gene sequence datasets

The sequence data was used in two ways. In the first case, we used the sequences of a total of 42 known motifs (Table 3.1) extracted by Pilpel (Pilpel et al., 2001). It was composed of 42 motifs (Table 1). We then scanned the upstream regions of ORFs for the presence of these motifs using the AlignACE program (Hughes et al., 2000). The sequence profile was represented by the occurrence of these motifs in the promoters of each gene in the genome.

In the second case, we analyzed the relationship between the expression profiles and the raw upstream sequences. We extracted gene upstream sequences $\sim$ 1kb from each gene. From these sequences, we calculated the frequency of all possible $l$-mers in each gene. For $l = 5$, each gene had 1024 ($= 4^5$) base combinations. The sequence profile was encoded in the frequency of $l$-mers.

We applied the kernel as $k(\mathbf{x}_{seq}, \mathbf{x'}_{seq}) = (\mathbf{x}_{seq}^T \mathbf{x'}_{seq})^d$ to the sequence data. When $d = 1$, it is the linear kernel, and when $d > 1$, it is the polynomial kernel.

### 3.2.4 Measurement of the effect of motif combinations

To measure the effect of the motif pairs, we defined the ECRScore (Expression Coherence coRrelation Score) calculated by a Pearson correlation coefficient of expression profiles for all possible pairs of genes whose upstream regions had the two motifs, $m_i$ and $m_j$:

$$ECRScore(m_i, m_j) = \frac{N_\tau(m_i \bigcap m_j)}{N(m_i \bigcap m_j)}, \tag{3.7}$$

where $N(m_i \bigcap m_j)$ is the number of all pairs of genes whose upstream regions have the two motifs, and $N_\tau(m_i \bigcap m_j)$ is the number of gene pairs whose correlation coefficient is larger than the threshold $\tau$. The threshold was chosen based on the fifth percentile of the distribution for correlation coefficients of randomly sampled gene pairs.

## 3.3 Results

We applied a computational method, kernel CCA, to the identification of novel transcriptional regulatory elements. The main purpose of our experiments was to find regulatory motifs that were associated with gene regulation in specific biological processes. Using the kernel CCA, we first found highly correlated features between expression profiles and the sequence motifs. The key motifs in gene regulation were then identified from the weight scheme by the kernel CCA (see Methods section). Furthermore we demonstrate that it is possible for our method to be applied for identification of motif pairs using raw upstream sequences.

### 3.3.1 Identification of the relationship between gene expression and known motifs

We first explored the relationship between gene expression profiles and known motifs using a yeast gene expression dataset related to the cell cycle (Spellman et al., 1998) and a set of known motifs (see Table 3.1) extracted by AlignACE (Pilpel et al., 2001). A total of 551 ORFs (open reading frames) in the expression dataset contained at least one known motif. In the parameter setting, the degree of polynomial kernel was set to 3, the parameter $\sigma$ in Gaussian RBF kernel was 0.5, and the regularization parameter was 0.1. These parameters were chosen based on the parameter setting that produced a high correlation from multiple runs.

The results from the kernel CCA were visualized using the CC1 (first canonical correlation) score (Figure 3.2). In Figure 3.2, each point corresponds to a gene, and a cloud of the diagonal points illustrated the correlation between the expression and the motifs. The shape of diagonal points and the high correlation coefficient (0.996) indicated that the kernel CCA was able to find the close relationship between the expression profiles and the sequence motifs. We then performed the linear canonical correlation analysis using the same datasets. The correlation coefficient (0.612) obtained from the linear CCA was much lower. As shown in Figure 3.3, the linear CCA could not identify the significant correlation between expression profiles and motifs. This further supports that kernel CCA improve significantly in finding the correlation between the two datasets.

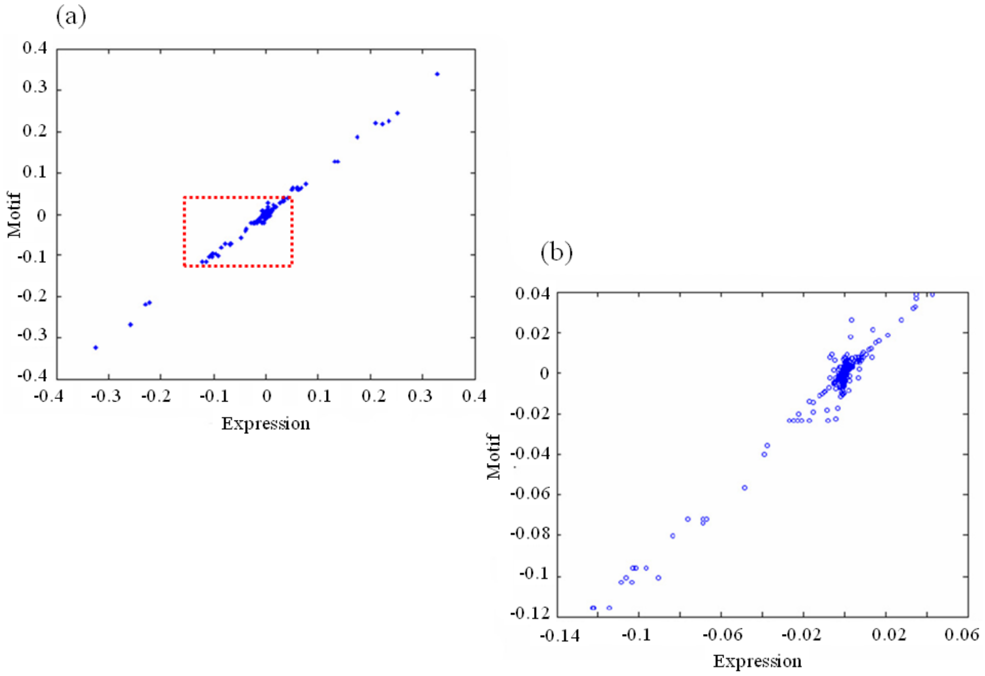


Figure 3.2: Relationship between gene expression profiles and regulatory sequence motifs. (a) The plot shows the correlation between gene expression profiles and the regulatory sequence motifs. Each dot represents one gene in the dataset, and x-axis means the value of $u_{exp}$, y-axis is $u_{seq}$. (b) The plot is a close-up view of the boxed area in (a).

The motifs were searched by the weight function of Equation 3.6 (see Methods section) with the model obtained by the kernel CCA and the top ranked motifs are shown in 3.2. SWI5 motif, a binding site of SWI5 protein, has the highest weight value. SWI5 has been known to act in G1 phase and in the M/G1 boundary in the cell cycle (Dohrmann et al., 1992, 1996). SFF' motif is a binding site of FKH1

Table 3.2: The list of top ranked motifs based on the weight scheme by the kernel CCA.

| Motif | Weight | Function |
|---|---|---|
| SWI5 | 0.89026 | Transcription Activation in G1 phase |
| SFF' | 0.45399 | FKH1 binding site that regulate the cell cycle |
| MCB | 0.29633 | MBF binding site that activates in late G1 phase |
| LYS14 | 0.21796 | Lysine biosysthesis pathway |
| ALPHA2 | 0.16532 | Encoding a homeobox-domain |

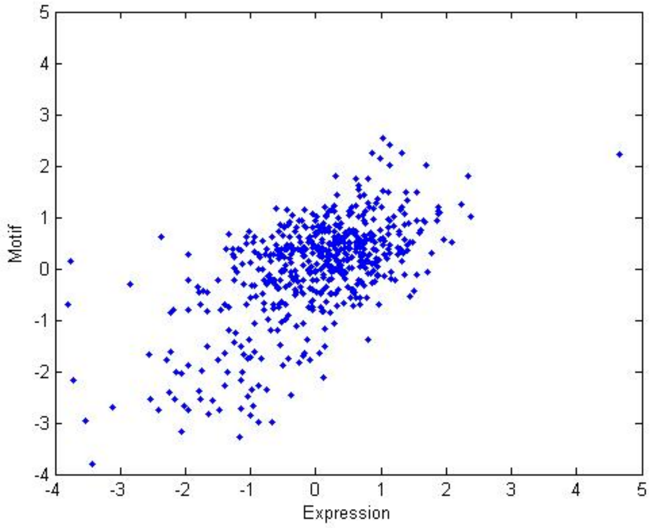


Figure 3.3: Relationship between gene expression profiles and regulatory motifs from the linear CCA

transcription factor that affects the expression of genes controlling the cell cycle during the G2-S phase change (Morillon et al., 2003). The MCB motif is one of the well-known motifs in the yeast cell cycle as a binding site in the MBF protein complex. MBF protein is composed of MBP1 and SWI6, and MBP1 is a DNA binding component while SWI6 has regulatory roles. It is well known that the MBF protein complex regulates the transcription of many genes in the late G1 phase (Dohrmann et al., 1992; Simon et al., 2001). ALPHA2 protein also plays a role in the cell cycle. It operates synergistically with MCM1 protein to repress the expression of its target genes (Vershon and Johnson, 1993; Zhong et al., 1999). MCM1 protein is a key regulator involved in the transcription of several M/G1 genes during the cell cycle (Simon et al., 2001; Lydall et al., 1991). A high weight value of ALPHA2 is supported by the evidence that ALPHA2 protein binds to the MCM1 protein and influences the regulation of other cell cycle-related genes (Keleher et al., 1989; Mead et al., 1996). Using the set of known motifs, our results are consistent with previous reports, validating the analysis method employed.

To further validate the result of top-ranked motifs extracted by kernel CCA, we compared the weights obtained from cell cycle-related ORF set with those obtained from randomly selected set. We performed the same procedure using random ORFs that are not known to be related to the cell cycle. Figure 3.4 shows the highly weighted motifs obtained from our method in cell cycle-related gene set and non cell cycle set, and the relative positions of those motifs are presented in the weight distribution of all motifs. The weight values obtained from random set were significantly lower than those obtained from cell cycle-related ORF set. We could infer that the significantly correlated motifs were not extracted from these random datasets. In summary, our method could identify the regulatory motifs that have high weights indicating high correlation between the upstream sequences and the gene expression profiles.

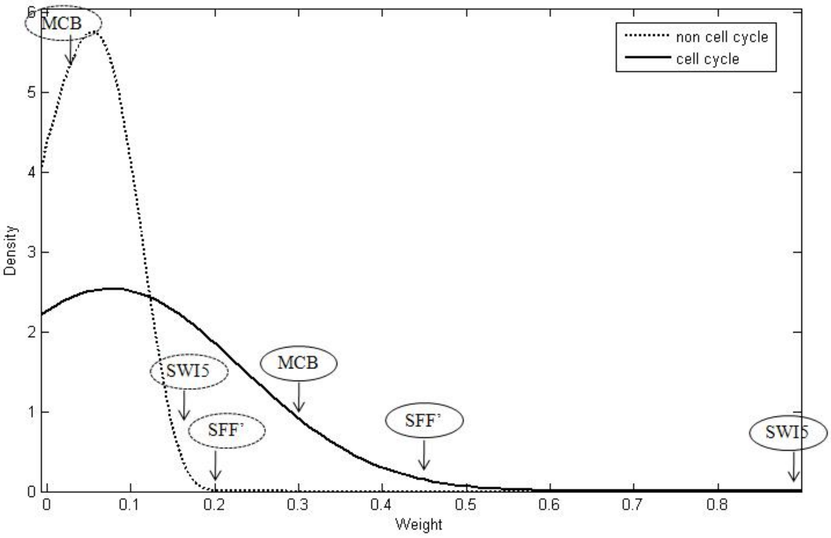


Figure 3.4: Weight distributions for MCB, SFF' and SWI5 motifs derived from cell cycle and non cell cycle-related datasets The dotted line indicates the weight distribution from the non-cell cycle datasets and the solid line from cell cycle datasets.

### 3.3.2 Identification of cell cycle-related motifs

We then applied the linear kernel to the motif sequence data containing a total of 1,024 features (window size $l = 5$) extracted from the raw upstream sequences of genes and Gaussian RBF kernels with parameter $\sigma$ values of 0.3 to the expression data. The regularization parameter was set to 0.1. These parameters are also empirically chosen based on the fact that they produced a high correlation. Figure 3.5 shows the CC1 score which represents the correlation between the expression profiles and the sequence patterns. When the linear kernel was applied to the sequence dataset, the expression data is closely related to the motif data using the

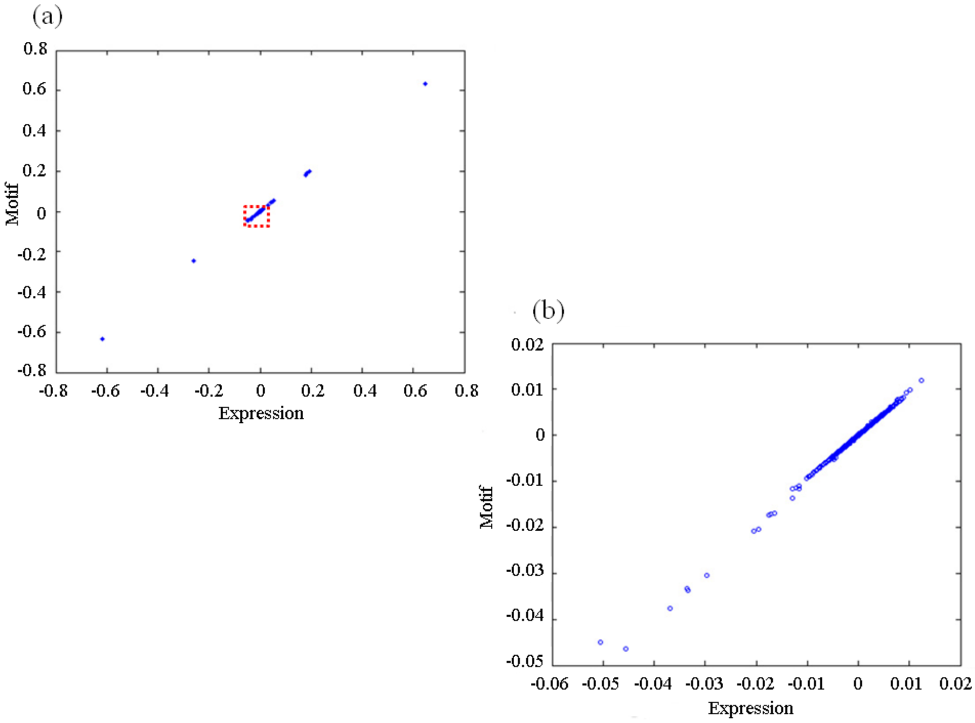


Figure 3.5: Correlation between expression profiles and motifs derived by using the raw upstream sequence data. The plot on (b) is an enlargement of the boxed area in (a).

raw sequences of 5-mers.

The 5-mer motif patterns with high weights are listed in Table 3.3. The 5-mer with the highest weight is 5'-GCGTG-3', which is similar to the MCB motif (5'-ACGCGT-3'). As described previously, MCB is an important motif involved in the cell cycle. The second-ranked sequence (5'-CGTGT-3') matched to the first five bases of the ALPHA2 motif sequence. From the second component, we also found several significant sequences, including a consensus sequence (5'-CGCGT-3') that

Table 3.3: High-scored motifs in the first and the second components using 5-mer raw upstream sequences.

| Sequence | Motif Description | Weight | Component | Rank |
|---|---|---|---|---|
| GCGTG | MCB (ACGCGT) | 0.079567 | 1 | 1 |
| CGTGT | MATalpha2 (CRTGTWWWW) | 0.075340 | 1 | 2 |
| CATGT | MATalpha2 (CRTGTWWWW) | 0.046299 | 1 | 12 |
| CCGGA | MCM1 (CCNNNWWRGG) | 0.044133 | 1 | 13 |
| TAAGG | MCM1 (CCNNNWWRGG) | 0.042387 | 1 | 15 |
| CCACG | SCB (CACGAAA) | 0.018992 | 2 | 4 |
| CGCGT | MCB (ACGCGT) | 0.017870 | 2 | 5 |
| GTGTT | MATalpha2 (CRTGTWWWW) | 0.016595 | 2 | 9 |

is identical to the MCB motif (5'-ACGCGT-3'). This further confirmed that the MCB motif affects gene expression in the cell cycle. Another interesting motif is 5'-CCACG-3', which is a sequence block with one base shift from the known SCB motif (5'-CACGAAA-3'). The SCB motif is a binding site of the SBF protein, which is a complex of SWI4 (a DNA-binding component) and SWI6 (a regulatory component) (Simon et al., 2001), and SBF is a major regulator in the G1/S transition.

### 3.3.3 Combinational effects of regulatory motifs

We searched the motif pairs that have synergistic or co-regulatory combination effects in the yeast cell cycle. The regulatory mechanisms of eukaryotes are highly complex since most genes are normally synergistically regulated by different transcription factors. Therefore, identifying the synergistic motif combinations can contribute to systematically understanding the regulatory circuit.

In the present study, using the kernel CCA we calculated the weight value for each motif pair of 42 known motifs. The heat map of weight values of all motif pairs is provided in Figure 3.6. Table 3.4 presents the top ten motif pairs with the highest weight values and with occurrence of more than ten in all the investigated upstream sequences. It also shows ECRScores which represent gene expression coherence. All these scores are relatively high compared to the previously identified synergistic motif pairs (ECRScores $> 0.075$). As shown in Table 3.4, the pair with the highest weight value is MCB-MCM1. According to a previous study, MCB and MCM1 were characterized as a significantly cooperative motif pair in the regulation of the cell cycle (Das et al., 2004). Other highly ranked pairs, such as ECB-ALPHA2 and MCM1-ALPHA2, are already known that they are required for transcriptional regulation of early cell cycle genes. MCM1 activates transcription of ECB (early cell cycle box)-dependent genes during M/G1 phase (MacKay et al., 2001), and the MCM1 protein can interact with the ALPHA2 factor regulating the expression of mating-type-specific genes (Keleher et al., 1989; Mead et al., 1996). These evidences support that two ALPHA2-related motif pairs act synergistically in the expressional regulation of the yeast cell cycle process. The REB1 motif, a binding site of REB1 protein, is frequently found among the pairs of motifs with the highest weights. The REB1 protein is an RNA polymerase I enhancer-binding protein and binds to genes transcribed by both RNA polymerase I and RNA polymerase II (Morrow et al., 1989). It is a general regulator rather than a condition specific one. Therefore, it is reasonable that this protein shows a high frequency in our results. REB1-SWI5, REB1-MCM1' and REB1-ALPHA1 motif pairs are already identified as acting synergistically in the yeast cell cycle regulation (Banerjee and Zhang, 2003; Tsai et al., 2005; Hvidsten et al., 2005). Most of our results are consistent with the previous reports. In addition, it's worth noting that several previously uncharacterized motif

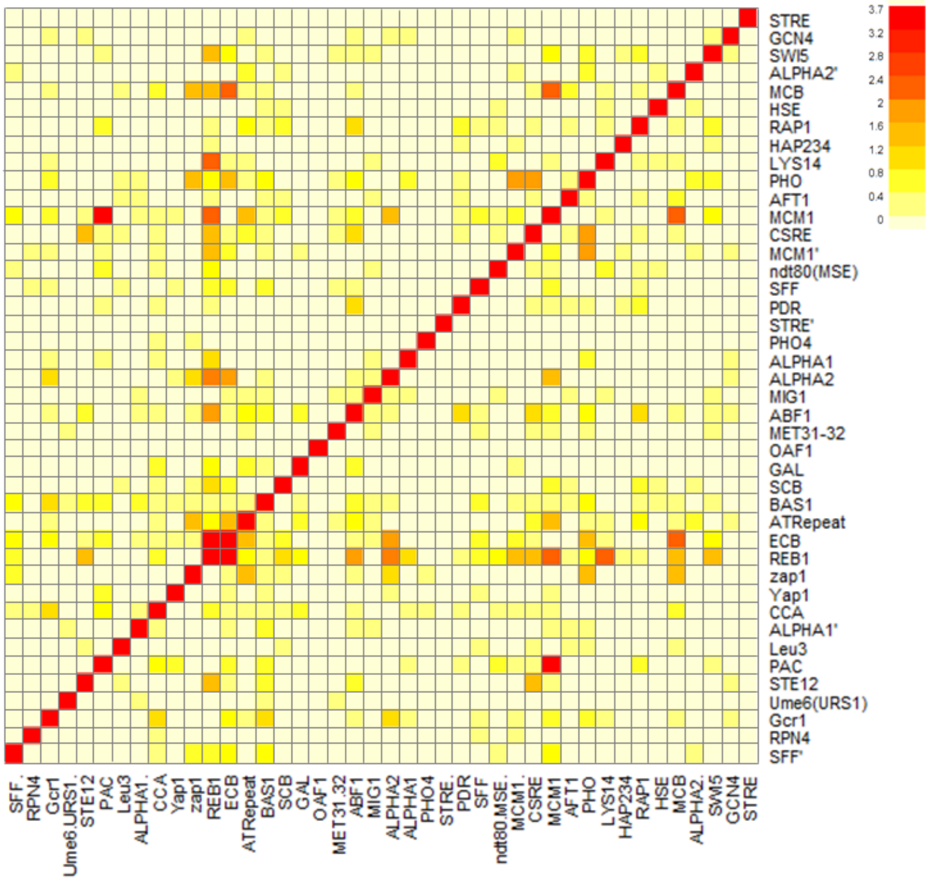


Figure 3.6: Heat map of weight values of motif pairs related to cell cycle regulation. Dark colors represent motif combinations of high weight values.

pairs were identified by our kernel CCA methods.

Table 3.4: The top 10 ranked motif pairs were extracted from the analysis of motif combination.

| Weight | Motif Pair | | ECRScore | Num. of ORFs |
|---|---|---|---|---|
| 2.5368 | MCB | MCM1 | 0.390 | 15 |
| 2.5018 | MCB | ECB | 0.439 | 12 |
| 2.0177 | PHO | MCM1' | 0.088 | 17 |
| 1.848 | ECB | ALPHA2 | 0.088 | 14 |
| 1.7535 | MCM1 | ALPHA2 | 0.074 | 17 |
| 1.7263 | ATRepeat | MCM1 | 0.076 | 12 |
| 1.6995 | PHO | ECB | 0.127 | 11 |
| 1.6823 | REB1 | SWI5 | 0.099 | 14 |
| 1.6476 | REB1 | MCM1' | 0.115 | 13 |
| 1.4256 | REB1 | ALPHA1 | 0.067 | 15 |

## 3.4 Discussion

We presented a novel method that can identify the candidate conditional specific regulatory motifs by employing kernel-based methods. The application of the kernel CCA enables us to detect correlations between heterogeneous datasets, consisting of upstream sequences and expression profiles. From a data-mining perspective, our work is regarded as a new approach for detecting important features from regulatory sequences and gene expression profiles. We demonstrated that major motifs in a specific biological process can be extracted by a CC score via modelling a close relationship between two datasets related to gene regulation.

As genome-wide datasets of various types become available, it's important to analyze these datasets in an integrated manner (Kasturi and Acharya, 2005). It is possible to come up with novel biological hypotheses by integrating diverse biological resources generated for specific research purposes. In these aspects, the kernel CCA is regarded as a useful method that can extract the biological factors with significant roles by integrating different types of biological data. Many studies for identifying motifs have been based on sequence conservation or sequence characteristics, regardless of the biological processes. Therefore our method can be regarded as complementary approach in the analysis of gene regulation.

Our method found important motifs related to the cell cycle by using raw upstream sequences as well as known motif sets. In the present study we used the raw sequences of window size, $l$=5. If we enlarged the window size, the dimension for sequence features increased exponentially, whereas the frequency of motifs decreased. Although the window size used in our experiments was shorter than the length of several known transcription factor binding sequences, it was long enough to obtain worthwhile results.

In the future research, we will apply the proposed method to diverse gene expression datasets, especially cancer-related datasets. The cancer-related regulatory program can be elucidated by analyzing regulatory motifs from a set of enriched genes in the cancer transcriptome (Rhodes et al., 2005). Using the kernel CCA, a correlation analysis between regulatory sequences and the cancer transcriptome may directly catch regulatory motifs related to the abnormal gene regulatory program.

# Chapter 4

# Investigation of combinatorial effects of DNA methylation

## 4.1 Background

The addition of a methyl group to cytosine residues in the context of CpG dinucleotides (i.e., 5-methylcytosine) by the DNA methyltransferease (DNMT) enzymes is the most well studied epigenetic event. DNA methylation is known to play significant roles in many cellular processes, including embryonic development, genomic imprinting, X-chromosome inactivation, and preservation of chromosome stability. In addition, aberrant DNA methylation has been shown to disrupt many cellular processes and is frequently observed in most human diseases, including cancer (Suzuki and Bird, 2008; Robertson, 2005; Esteller, 2008; Keshet et al., 2006).

Methylation in CpG islands (CGIs), particularly in the promoter and first exon regions, is known to block genomic binding sites of activating transcription factors or other proteins and it is strongly associated with gene repression (Suzuki and Bird, 2008; Jones and Takai, 2001). In particular, the effect of DNA methylation on tumor suppressor genes (TSGs) has been extensively studied (Ueki et al., 2002). Transcriptional silencing of this key class of genes could contribute to defective regulatory processes in cancer, and the promoter CGI hypermethylation of TSG has been observed in a various types of cancers (Sakai et al., 1991; Merlo et al., 1995). However, few studies have examined the complex relationship between DNA methylation and gene expression on a genome-wide scale using accurate, high-resolution DNA methylation data.

Profiling of methylated CpG sequences is now possible by using next generation sequencing technologies and a number of recent studies have used high-throughput approaches to study DNA methylation (Chavez et al., 2010; Kim et al., 2011). Although generating enormous amounts (terabytes) of data is possible, single-base pair resolution of bisulfite-converted DNA is still costly and highly labor intensive. Recently, cost effective, genome-wide methylation approaches that do not rely on bisulfite-treated DNA have been developed, including methylation-sensitive restriction enzymes approaches (Zuo et al., 2009). One approach, the methylated-CpG island recovery assay (MIRA) (Rauch and Pfeifer, 2010) followed by sequencing (mCpG-seq), utilizes methylated-CpG-binding protein complexes with high affinity to methylated CpG dinucleotides in genomic DNA. Now a technique known as MBDCap-seq (Brinkman et al., 2010) is able to utilize double-stranded DNA, does not depend on the application of methylation-sensitive restriction enzymes, and generates DNA sequence variation data (Robinson et al., 2010).

The availability of high resolution DNA methylation and gene expression data on a genome scale now allows scientists to investigate the functional consequence of DNA methylation in various genomic regions, including CGIs which have been extensively investigated in the literature (Esteller, 2007; Bell et al., 2011; Pai et al., 2011). CGIs are often found near the promoter regions of genes and the CGI hypermethylation is known to have significant inhibitory effect on gene expression. In normal cells, CGIs are protected from methylation. However, hypermethylation of promoter CGIs of important genes, i.e. TSGs, is frequently observed in cancer cells (Sproul et al., 2011). In addition to CGIs, recent studies have reported that DNA methylation of other genomic regions can alter downstream gene expression. It was recently reported that methylation of CGIs near transcription start sites (TSSs) of genes (Sproul et al., 2011) or in CGI shores (Irizarry et al., 2009), regions about 2kb outside of CGIs, were both strongly associated with gene expression. In addition, a strong correlation between methylation in the first exon and expression of the corresponding genes was demonstrated (Brenet et al., 2011). Although these recent studies have clearly shown an association between DNA methylation at various genomic regions and gene expression, several questions remain to be answered. Specifically, in our study on the breast cancer cells, research questions are: How does DNA methylation in the different genomic regions contribute to gene expression? Are there subtype specific DNA methylation-gene expression patterns in breast cancer? Does the methylation of transcription factor binding sites impact transcription factor binding and subsequent gene expression?

To answer these questions, we used genome-wide profiling data from 30 breast cancer cell lines from the Integrated Cancer Biology Program (ICBP, http://icbp.nci.nih.gov/). We integrated MBDCap-seq methylation data and Affymetrix microarray gene expression data (Neve et al., 2006). The important goals of our study were:

1. Genomic studies have established major breast cancer intrinsic subtypes that show significant differences in incidence, survival and response to therapy (Koboldt et al., 2012). Basal-like breast tumors display aggressive clinical behavior and belong to the high-risk breast cancers that typically carry the poorest prognoses (Fadare and Tavassoli, 2008; Toft and Cryns, 2011). To

investigate whether phenotype specific methylation and expression patterns exist in the basal A, basal B, and luminal breast cancer molecular subtypes, we used an information-theoretic approach to identify genes with differentially methylated DNA regions and differential expression levels.

2. To perform an integrated analysis of DNA methylation and gene expression data on a genome-wide scale and to detect subtype-specific effects of DNA methylation in breast cancer cells. We examined relationships between DNA methylation and gene expression using step-wise analysis starting from genes whose expression was significantly altered in a particular subtype.

3. We used Pearson's correlation analysis and decision tree learning to investigate the effect of DNA methylation in various regions (CGIs, CGI shores, promoter regions, 1st exons, 1st introns, and 2nd exons) on the breast cancer subtype differential gene expression.

4. To investigate relationship between transcription factors and DNA methylation in promoter regions, we examined the relationship between DNA methylation specifically at transcription factor binding sites (TFBSs) and gene expression in the breast cancer molecular subtypes.

## 4.2 Materials and methods

### 4.2.1 Data

We prepared methylation and gene expression data from 30 breast cancer cell lines representing three tumor phenotypes found in patients (Neve et al., 2006): basal A, basal B, and luminal subtypes. Among 30 cell lines, 17 were basal-like and 13 were luminal-like subtypes (Table 4.1). The basal-like 17 cell lines were further subdivided into 7 basal A and 10 basal B subtypes.

Gene expression data from Affymetrix microarray experiments (Neve et al., 2006) was downloaded. Genome-wide methylation profiles were measured using the MBDcap-seq technique. The double stranded methylated fragments were sequenced and reads were mapped to the human reference genome. Methylation levels were calculated by measuring the density of the read coverage (Rao et al., 2013), as we have described previously.

The microarray gene expression data were processed and analyzed using R and Bioconductor. The expression values were centered by mean-adjusting each log abundance value (subtracting each value from the mean expression value in the cell line).

### 4.2.2 Profiling of DNA methylation patterns

To investigate DNA methylation characteristics across the 30 breast cancer cell genomes, methylation profiles were measured on ± 10 kb genomic regions around the TSS of each gene. We divided the genomic regions into bins with a size of 100 bases. DNA methylation levels were then measured as the number of mapped reads within each bin.

### 4.2.3 Identifying differentially methylated/expressed genes by information theoretic analysis

We identified differentially methylated and expressed genes in the three breast cancer subtypes using normalized entropy. Entropy is a measure of uncertainty, defined as follows:

$$\mathrm{H} = -\sum_{i=1}^{n} p_i \log p_i$$

Table 4.1: 30 Breast cancer cell lines and molecular subtypes

| Cell line | Subtype | Cell line | Subtype |
|---|---|---|---|
| BT549 | BaB | HCC1569 | BaA |
| HCC1937 | BaA | HCC1143 | BaA |
| HCC1428 | Lu | HCC202 | Lu |
| MDAMB436 | BaB | SUM185PE | Lu |
| 600MPE | Lu | HCC1500 | BaB |
| MDAMB231 | BaB | SUM225CWN | BaA |
| SKBR3 | Lu | MDAMB453 | Lu |
| SUM1315MO2 | BaB | SUM52PE | Lu |
| HSS78T | BaB | MCF12A | BaB |
| MDAMB157VII | Lu | HCC70 | BaA |
| HCC1954 | BaA | SUM149PT | BaB |
| GCC2185 | Lu | LY2 | Lu |
| MCF7 | Lu | BT20 | BaA |
| MCF10A | BaB | BT474 | Lu |
| SUM159PT | BaB | AU565 | Lu |

Lu: luminal; BaA: basal A; BaB: basal B

where $p_i$ denotes the probability of the state $i$, and $n$ is the total number of the states. In this study, the state $i$ is a cancer phenotype, i.e. $i = (basalA, basalB, Lu)$. For methylation profiles, the probability $p_i$ is measured by $t_{ji}/c_j$, where $c_j$ is sum of read counts for cell lines in a genomic region $j$ and $t_{ji}$ is sum of reads for a phenotype $i$ in the region $j$. For gene expression, $c_j$ is sum of expression values for cell lines in a gene $j$ and $t_{ji}$ is sum of expression for a phenotype $i$ in the gene $j$. The entropy $H$

achieves its maximum value when all states are equally probable, that is, it exhibits the lowest degree of uncertainty. If there is only one state, then the entropy $H$ is zero.

Normalized entropy is the ratio of entropy to maximum entropy as follows:

$$\mathrm{H}_0\left(x\right) = \mathrm{H}\left(x\right) / \mathrm{H}_{\max}$$

where $H_{max}$ is maximum entropy value where the probabilities are all equal.

We measured the normalized entropy and identified differentially methylated regions and differentially expressed genes. To avoid errors on the probability calculation, we introduced pseudo-probability to every zero-valued position.

### 4.2.4 Identifying downregulated genes in each subtype for integrative analysis

Genes differentially expressed in each different molecular subtype were further identified as follows. Suppose that $e_{gl}$ is an expression level of a gene $g$ in a cell line $l$. Since the cell line $l$ is clustered into a specific subtype $i$, we calculate the median values $Median(e_g, i)$ for the expression levels in each subtype $i$ per gene $g$. In this study, we measured three median value $Median(e_g, Lu)$, $Median(e_g, BasalA)$, $Median(e_g, BasalB)$ for each gene $g$.

If the median value $Median(e_g, i)$ of a gene $g$ in a type $i$ was significantly lower than those of other two types, we defined the gene $g$ as down-regulated in a specific type. In our study, log-ratio 1.5 was the criterion for significance.

### 4.2.5 Correlation between DNA methylation and gene expression

To investigate the relationship between methylation in various regions and gene expression in the 30 breast cancer cells, we examined methylation levels in gene

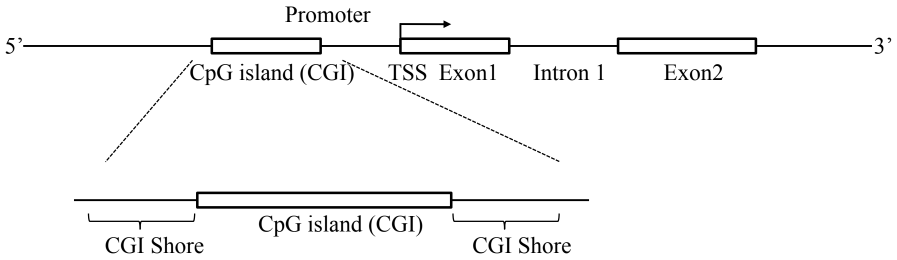


Figure 4.1: Genomic regions for studying DNA methylation profiles. A gene body is composed of promoter and coding regions including exons and introns. CGIs as well as these regions were studied for the effect of DNA methylation on gene regulation.

promoter regions (2kb upstream regions from TSSs), CGIs, CGI shores, the first and second exon and the first intron (Figure 4.1). The association between gene expression and methylation values of these datasets was measured by a Pearson's correlation coefficient. It was calculated on the paired data of a gene expression level and the methylation level in the genomic region.

### 4.2.6 Combinatorial effects of DNA methylation in various genomic regions

To identify which regions have dominant effects on downstream gene expression and also to investigate on the combinatorial roles of DNA methylation of the various genomic regions in each subtype, a decision tree was constructed using the methylation profiles in each region. For the learning purpose, a gene was an instance of data and gene expression was considered as a class variable, *i. e., up or down regulated genes*. The methylation value in each genomic region was an attribute. For binary classification, in training dataset of each subtype, the class values were discretized to high and low, *i. e., upregulated or downregulated genes*. If a gene was significantly downregulated in a subtype but the gene was upregulated in the other subtypes, the class values of the genes in the cell lines within the subtype were designated as low. For example, assume that the expression of a gene is significantly downregulated in Lu subtype. Then among 30 cell lines, 13 instances with Lu subtype are marked as low and 17 with the other types are high. The trees were built using REPTree in WEKA software (Hall et al., 2009).

### 4.2.7 Analysis of transcription factor binding regions possibly blocked by DNA methylation

For the integrative analysis of TFs, DNA methylation and gene expression, we used four datasets: gene expression, methylation profiles, cell specific DNA sequences and information for TF binding sites (TFBSs; TRANSFAC database (Matys et al., 2006)). We considered only downregulated genes in each subtype, as we were most interested in DNA methylation of TFBSs, possible interference on TF binding, and subsequent negative effect on gene expression. We referred to these downregulated as *target genes*. Differentially methylated genomic regions of the target genes were identified by statistical testing (t-test) of methylation levels at each 100bp-bin for the promotor regions. Cell-specific consensus sequences were computed by assembling short reads in the promotor regions of these genes. TFBSs were searched on the cell-specific consensus sequences corresponding to the hypermethylated bins, using ′minimize false positive′ option of the match tool in the TRANSFAC package (Kel et al., 2003).

Among the collected TFs that could be potentially blocked by TFBS methylation in the promotor region, we selected TFs whose expression levels were not significantly different in each phenotype (by t-test), as to exclude cases where the down-regulation of the target genes is as a result by difference in the expression levels of TF, an activator gene. In this way, we compiled cases where down-regulation of the target genes was due only to the hypermethylation in the promotor region, not other factors, such as the genomic sequences on the TFBSs and the expression levels of the TF.

## 4.3 Results

### 4.3.1 DNA methylation in 30 ICBP cell lines

We measured and compared the methylation density of 2kb promoter regions for all genes in 30 breast cancer cell lines. Figure 4.2 shows subtype-specific density plots of promoter regions, excluding unmethylated genes. Overall, the methylation density was similar in each subtype. We observe that the number of highly methylated ($>$ 50) promoter regions tended to be lower in BaB. The density of the regions whose methylation levels were over 50 was around 10% in Lu and BaA, but 4% in BaB.

Next, we investigated CGI methylation around each gene. CGIs are defined as regions of DNA of greater than 500 bp with a G+C equal to or greater than 55% and observed CpG/expected CpG of 0.65 (Takai and Jones, 2002). Using the position information of the CGIs from UCSC genome browser, we checked the methylation profile in the CGI near each gene. In the 30 breast cancer cell lines, approximately 47% of CGIs were methylated; however, distinct methylation density for each subtype was not apparent (Figure 4.3).

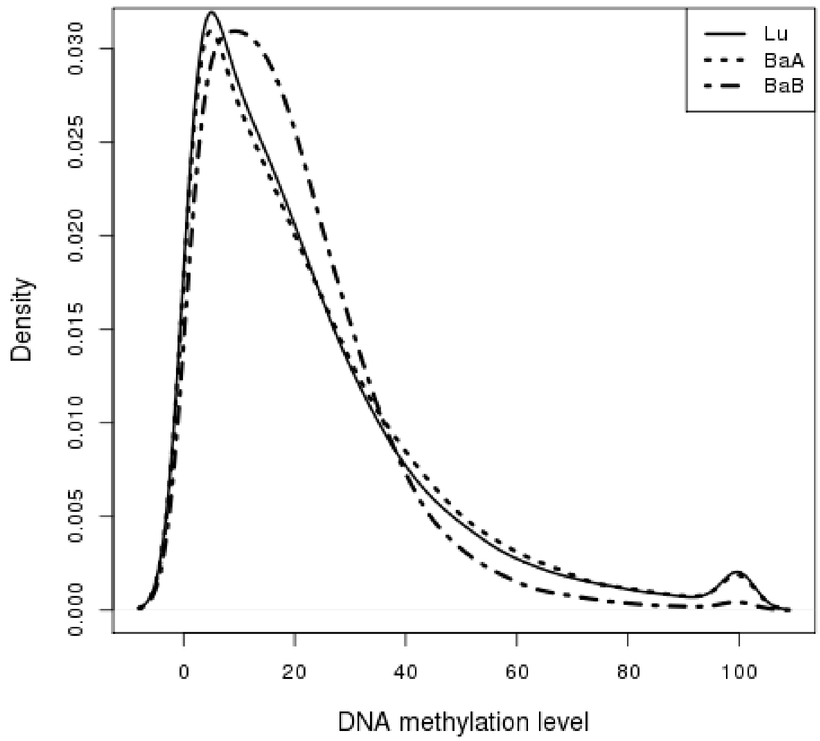


Figure 4.2: Methylation density of promoter regions in 30 breast cancer cell lines. Density was measured for each subtype. The methylation levels are on the x-axis and the y-axis is probabilistic density. Unusual bulbs around 100 on the x-axis were because methylation levels over 100 were truncated to 100. Lu, luminal; BaA, basal A; BaB, basal B.

### 4.3.2 Information theoretic analysis of phenotype-differentially methylated and expressed genes

To identify differentially methylated and expressed genes across the breast cancer genome, we measured normalized entropy. Lower entropy corresponded to genes

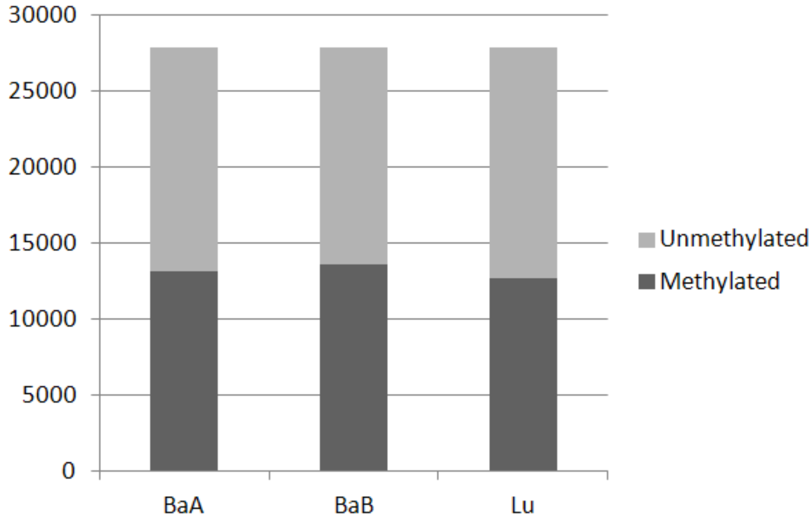


Figure 4.3: Average number of methylated and unmethylated CGIs in each cell. The unmethylated means that the mapped read count is zero in the CGI. BaA: basal A, BaB: basal B, Lu: luminal.

more differentially methylated or expressed in each subtype. First, we determined which genes were differentially methylated. Considering only genes with >3 mapped reads, there were 241 differentially methylated genes with the entropy threshold 0.2 and 564 differentially expressed genes with entropy threshold 0.5. Among these, only three genes were common to both the differentially methylated and expressed gene sets (Table 4.2) Thus, we concluded that separate analysis of differentially methylated and expressed gene sets based on information theory is not effective for the integrated analysis of methylation and gene expression, although these methods were effective to highlight genes and genomic regions that were different according to phenotypes.

Table 4.2: Genes that were both differentially methylated and expressed

| Gene Name | Description |
|---|---|
| PLA2G12A | phospholipase A2, group XIIA |
| FAT1 | FAT tumor suppressor homolog 1 |
| PARP8 | poly (ADP-ribose) polymerase family, member 8 |

### 4.3.3 Integrated analysis of DNA methylation and gene expression

To perform the integrated analysis of DNA methylation and gene expression, we used a two-step analysis process: (1) identify differentially expressed genes in each subtype, and (2) for each genomic region, test if there is a strong negative correlation between methylation level at the genomic region and the expression level of the gene.

To select differentially expressed genes in each subtype, we measured median values of expression levels for each of the three breast cancer phenotypes. If the median value of a gene in one subtype was significantly higher or lower than the median value in the other two subtypes, the gene was considered to be differentially expressed in a specific type. For such differentially expressed genes, variations of methylation levels were then investigated.

As DNA methylation is known to inhibit gene expression and an inverse correlation between the DNA methylation and gene expression has been shown to exist, we were most interested in a negative correlation between DNA methylation and gene expression for the integrated analysis. As an example, Caveolin 1, *CAV1*, represents a negative relationship between DNA methylation and gene expression (Figure 4.4). The *CAV1* gene has been shown by us and others to regulate breast tumor growth and metastasis and is overexpressed in basal-like subtypes (Sloan et al., 2004; Savage et al., 2007; Rao et al., 2013). *CAV1* expression levels were clearly different in each

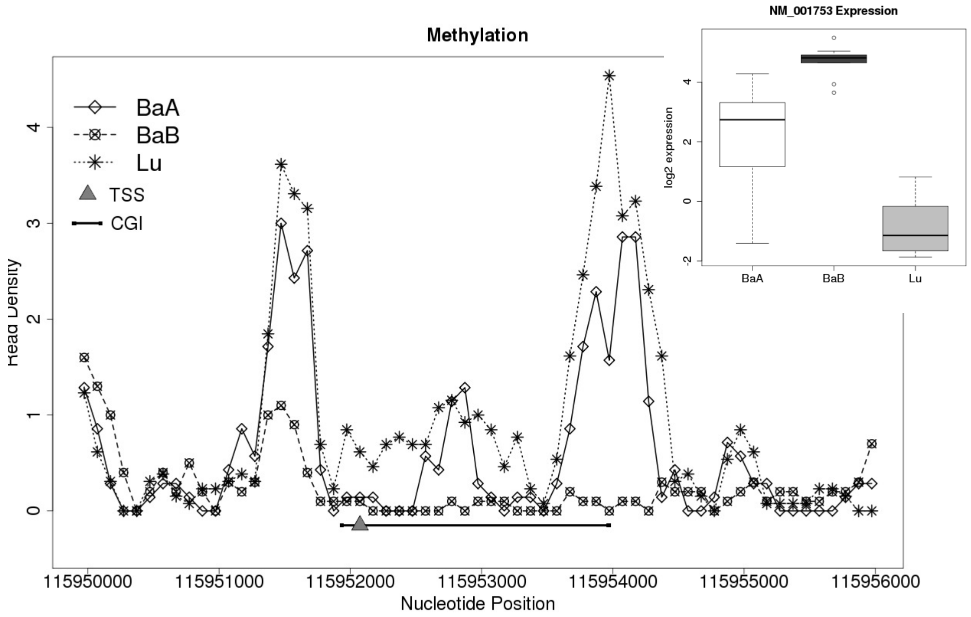
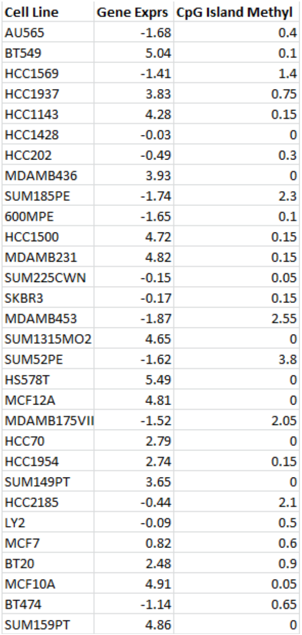


Figure 4.4: CGI methylation and gene expression of the CAV1 gene. Methylation and gene expression values from the 30 cell lines are grouped into luminal (Lu), basal A (BaA) and basal B (BaB) subtypes. **(a)** A plot showing the density of methylation in the CGI and shore regions located near the TSS of the CAV1 gene. The black bar shows the location of the CGI and the small orange triangle is the TSS. **(b)** A boxplot showing the expression of the CAV1 gene.

breast cancer subtype, higher in BaB subtypes and lower in Lu subtypes. However, when the DNA methylation profiles of the *CAV1* TSS and CGI were examined, methylation levels were significantly higher in the Lu compared to BaA and BaB. Furthermore, differential methylation of CGI shores, but not CGIs, significantly regulated *CAV1* expression, and breast cancer aggressiveness was associated with *CAV1*

(a)



| Cell Line | Gene Exprs | CpG Island Methyl |
|---|---|---|
| AU565 | -1.68 | 0.4 |
| BT549 | 5.04 | 0.1 |
| HCC1569 | -1.41 | 1.4 |
| HCC1937 | 3.83 | 0.75 |
| HCC1143 | 4.28 | 0.15 |
| HCC1428 | -0.03 | 0 |
| HCC202 | -0.49 | 0.3 |
| MDAMB436 | 3.93 | 0 |
| SUM185PE | -1.74 | 2.3 |
| 600MPE | -1.65 | 0.1 |
| HCC1500 | 4.72 | 0.15 |
| MDAMB231 | 4.82 | 0.15 |
| SUM225CWN | -0.15 | 0.05 |
| SKBR3 | -0.17 | 0.15 |
| MDAMB453 | -1.87 | 2.55 |
| SUM1315MO2 | 4.65 | 0 |
| SUM52PE | -1.62 | 3.8 |
| HS578T | 5.49 | 0 |
| MCF12A | 4.81 | 0 |
| MDAMB175VII | -1.52 | 2.05 |
| HCC70 | 2.79 | 0 |
| HCC1954 | 2.74 | 0.15 |
| SUM149PT | 3.65 | 0 |
| HCC2185 | -0.44 | 2.1 |
| LY2 | -0.09 | 0.5 |
| MCF7 | 0.82 | 0.6 |
| BT20 | 2.48 | 0.9 |
| MCF10A | 4.91 | 0.05 |
| BT474 | -1.14 | 0.65 |
| SUM159PT | 4.86 | 0 |

(b)

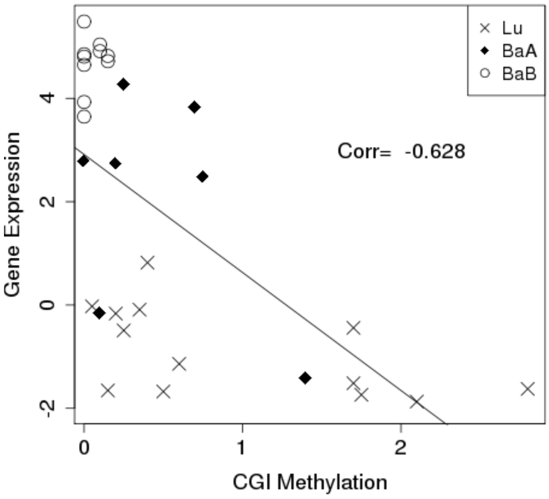


Figure 4.5: An example of the paired input data used to measure the Pearson correlation between gene expression and methylation. This paired data is for CAV1 gene. **(a)** Gene expression and CGI methylation across 30 cell lines. **(b)** Plot of gene expression profiles (y-axis) v.s. methylation levels (x-axis). Each pair in the cells is represented as a cross sign (Lu), a diamond (BaA) and a circle (BaB). A regression line is shown.

CGI shore methylation levels (Rao et al., 2013). The above negative correlation was measured by computing Pearson correlation coefficients. The Pearson correlation is measured by paired input data between DNA methylation profiles and gene expression levels across the 30 breast cancer cell lines. As an example, a correlation coefficient from CGI methylation and gene expression levels was calculated across

30 cell lines (Figure 4.5). The scatter plot for *CAV1* gene shows that gene expression and CGI methylation levels were negatively correlated.

We measured the methylation correlation for various genomic regions of downregulated genes in Lu and BaB subtype (Figures 4.6 - 4.7). Since only two genes were detected as downregulated in BaA subtype, the correlation results for BaA subtype were not included. Interestingly, when methylation in promoter regions was considered, several genes showed a clear negative correlation at the proximal regions of TSSs. Figure 4.6 is heatmaps that visualize promoter region methylation and downstream gene expression (light red colors mean that two vectors (methylation profiles and expression levels) were highly negatively correlated and bright green were positively correlated). In both Lu and BaB subtypes, strong negative correlations were observed in promotor regions, and methylation in the promotor regions near TSS showed strongest negative correlations. However, there were significant differences in promotor methylation patterns in Lu and BaB subtypes. In Lu subtypes, weaker negative correlations were observed at genomic regions further away from TSS. On the contrary, in BaB subtypes, consistently strong negative correlations were observed in entire promotor regions. This result implies that the DNA methylation on the promoter region has stronger epigenetic inactivation in Basal-like subtypes and the methylation of this regions may contribute to breast cancer progression.

Moreover, in most genes, first exon and CGI methylation levels were negatively correlated with expression levels (Figure 4.7). From the multi-exon genes, we measured correlation coefficients between the DNA methylation profiles for each exon and intron, and the expression level of the corresponding gene. A clear negative correlation was observed in the first exon, but this was not the case for second exons and first introns, a result consistent with a previous study showing that first exon methylation was closely associated with low gene expression (Brenet et al., 2011).

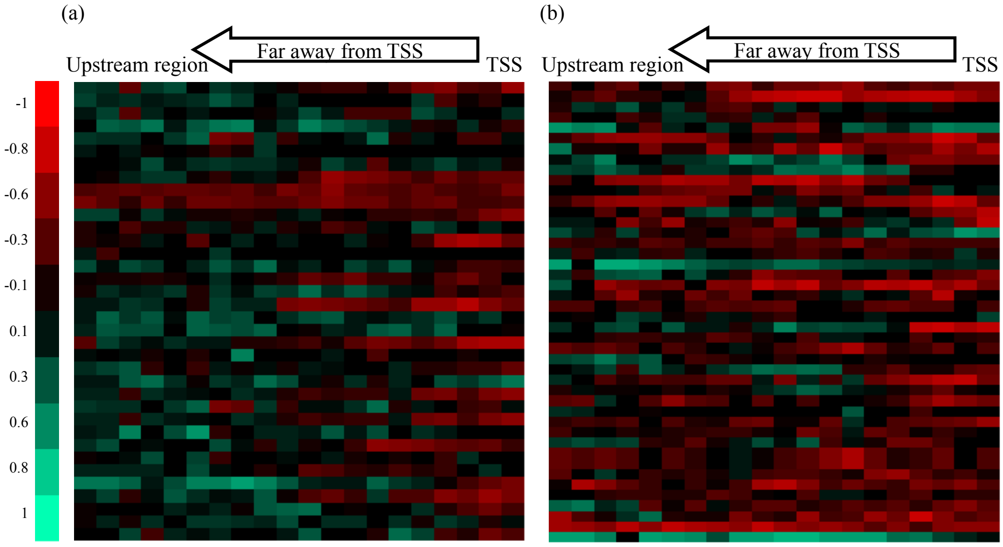


Figure 4.6: Correlation between promoter region methylation profiles and expression levels of genes downregulated in (a) Lu and (b) BaB subtypes. Unmethylated genes in the whole promoter region of 30 cell lines were excluded. Light red color was used for negative correlation and light green for positive correlation. Columns from right to left denote positions getting away from TSS. Each row is a downregulated gene in the subtype.

When we examined CGIs and CGI shore regions, negative patterns were also apparent. CGI and CGI shore DNA methylation levels were negatively correlated with gene expression levels in most genes, but in CGIs, much stronger relationships were shown in our datasets.

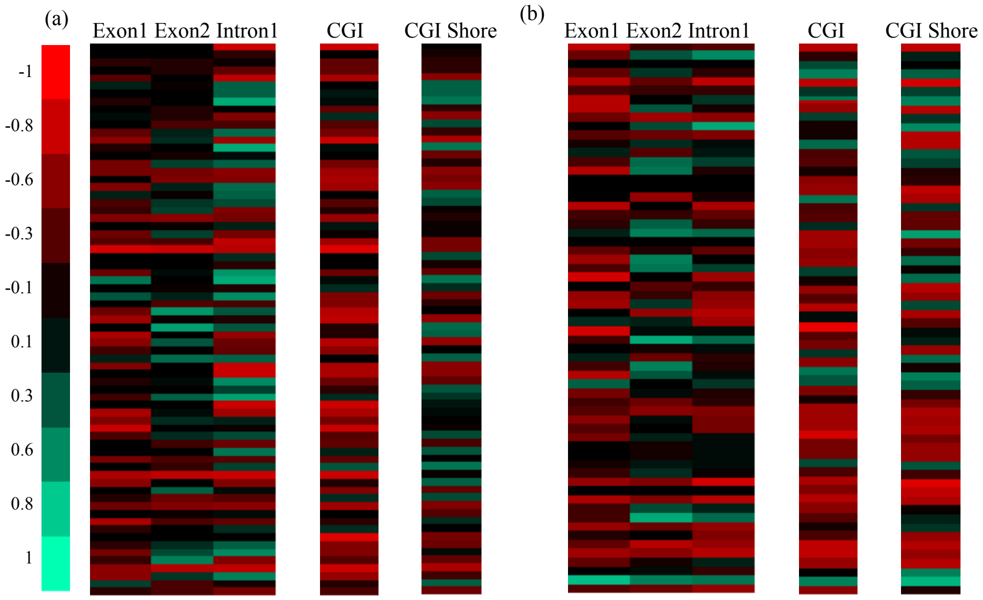


Figure 4.7: Correlation between methylation profiles on CGI, CGI shore, intron, and exon regions and expression levels of genes down-regulated in (a) Lu subtypes and (b) BaB subtypes.

### 4.3.4 Investigation of the combinatorial effects of DNA methylation in various regions on downstream gene expression levels

As DNA methylation occurs in many genomic regions, it was of interest to examine the effect of the various regions on downstream gene expression, particularly which regions may have a dominant effects on gene expression and whether the effects of the regions were similar in each subtype. Towards this goal, we performed a comprehensive study using six distinct genomic regions: promoter regions, CGIs, CGI shores, first and second exons, and first introns. Using the DNA methylation profiles in these regions, we performed a machine learning analysis.

The decision tree is a classification method that uses conjunctions of features for predicting target values in a tree-like hierarchical decision process. As decision tree learning identifies the most informative attributes for classification, this approach was used to discover regions with dominant and combinatorial effects on expression levels. We normalized the methylation levels of each region in a gene by adjusting the scale, then carried out the decision tree analysis.

The decision tree was constructed with a constraint of a maximum tree depth of three excluding leaf nodes, and in this case, the classification accuracy for genes, downregulated in Lu subtype, was 0.649 in a 10-fold cross validation (Figure 4.8 (a)). In the decision tree, the right-most branch means that the nodes in this branch were hypermethylated, and the left-most that the regions were hypomethylated. Consistent with the correlation analysis, CGIs were the most informative feature.

In the BaB subtype whose classification accuracy was 0.746 with the same maximum depth, the promoter regions and the first exons had combinatorial effects on gene expression (Figure 4.8 (b)). In the left branch of the decision tree where TSS1001-2000 were hypomethylated, it is intuitive that genes were unregulated. However, in the left branch, when TSS1-1000 was hypermethylated and also the first exons were hypermethylated, genes were down regulated. Note that TSS1001-2000 region had the dominant effect on the gene expression in the BaB subtype. This was consistent with our previous correlation analysis showing a clear negative correlation in much broader regions (Figure 4.6). Since CGI overlaps the first exon or promoter regions, we carried out the analysis again by separating into two cases: (1) CGI overlaps with the regions and (2) CGI does not overlap with the regions. Even when we separated CGI overlapping cases, the dominant factors (CGI for the Lu subtype and TSS1001-2000 for the BaB subtype) remained the same as when

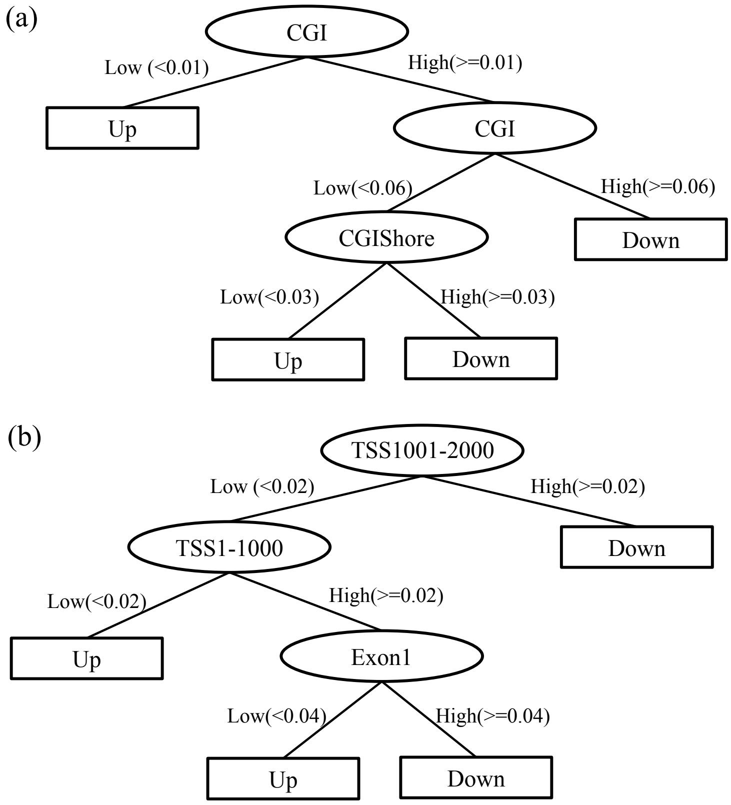


Figure 4.8: Decision tree analysis with downregulated genes in **(a)** Lu subtypes and **(b)** in BaB. The attributes are represented by circles, in where Exon1 is the first exon and CGIShore means 2kb outside region from CGI. TSS1-1000 means 1 to 1000 bp upstream region from TSS and TSS1001-2000 means 1001 to 2000 bp upstream. The Down in leaf nodes (rectangular boxes) means the gene is downregulated and Up means upregulated.

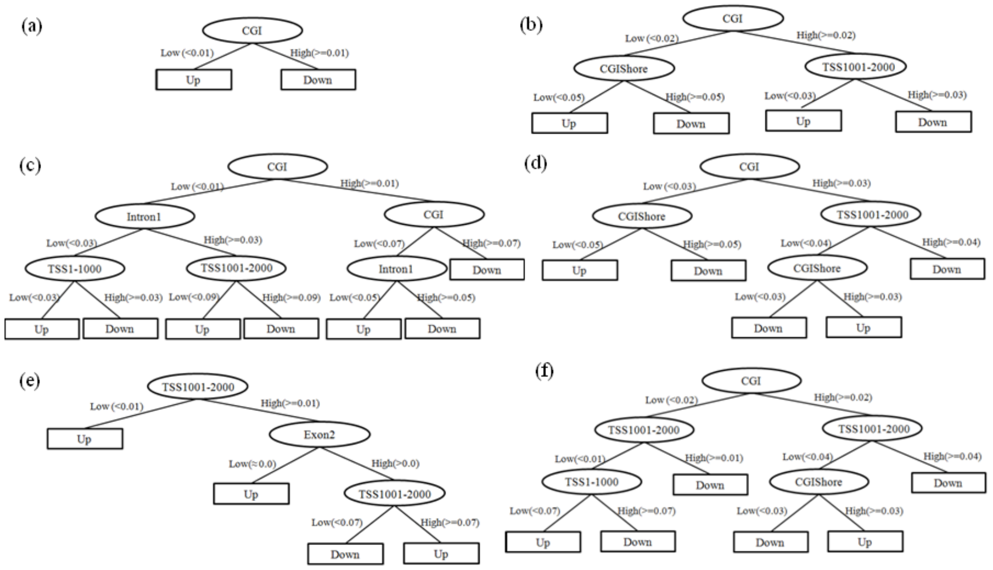


Figure 4.9: In case of down-regulation in Lu subtype, decision tree analysis separated by genomic regions of CGI. (a) Overlap with the first exon (The classification accuracy, Acc. is 0.737), (b) Nonoverlap with the first exon (Acc. is 0.590),(c) Overlap with TSS1-1000 (Acc. is 0.687), (d) Nonoverlap with TSS1-1000 (Acc. is 0.644), (e) Overlap with TSS1001-2000 (Acc. is 0.644) and (f) Nonoverlap with TSS1001-2000 (Acc. is 0.644).

we did not separate CGI overlapping cases. The decision trees when we did not separate CGI overlapping cases were presented in the main text (Figure 4.8) and the decision trees when we separated CGI overlapping cases were presented in Figures 4.9 and 4.10. The decision tree results suggest that altered gene expression in the two subtypes is associated with not only different promoter methylation profiles but also different combinatorial effects in various genomic regions.

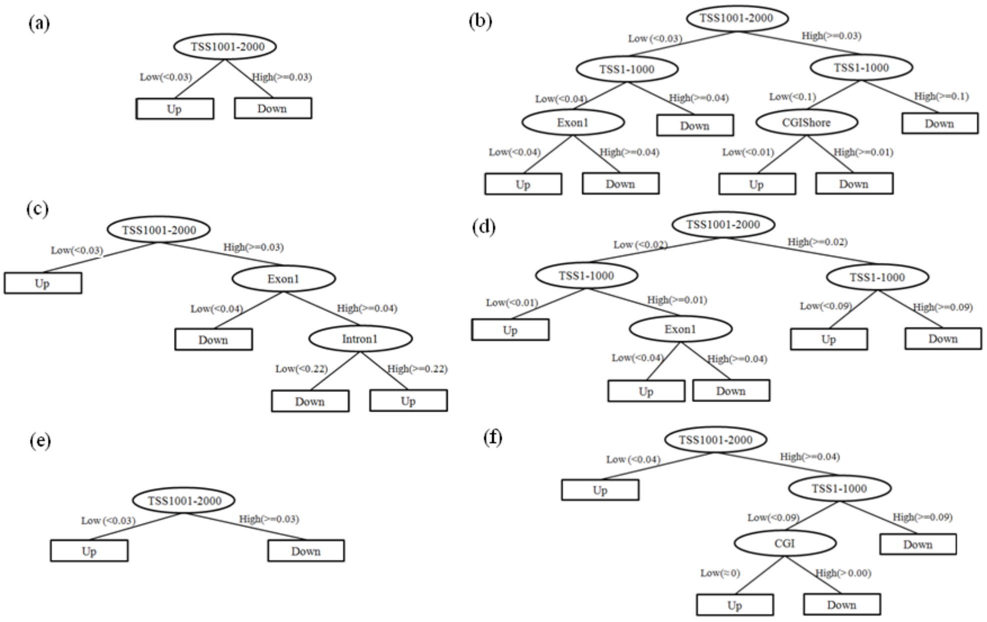


Figure 4.10: In case of down-regulation in BaB subtype, decision tree analysis separated by genomic regions of CGI. (a) Overlap with the first exon (Acc. is 0.773), (b) Nonoverlap with the first exon (Acc. is 0.760),(c) Overlap with TSS1-1000 (Acc. is 0.810), (d) Nonoverlap with TSS1-1000 (Acc. is 0.708), (e) Overlap with TSS1001-2000 (Acc. is 0.824) and (f) Nonoverlap with TSS1001-2000 (Acc. is 0.741).

### 4.3.5 Integrative analysis of transcription factors, DNA methylation and gene expression

We next sought to investigate the effect of DNA methylation on the interaction between TF and DNA, *i.e.* binding of a TF to the promotor region of a gene. To investigate this important concept, we developed a rigorous data mining protocol to compile a list of TF that are potentially blocked by DNA methylation. The schematic

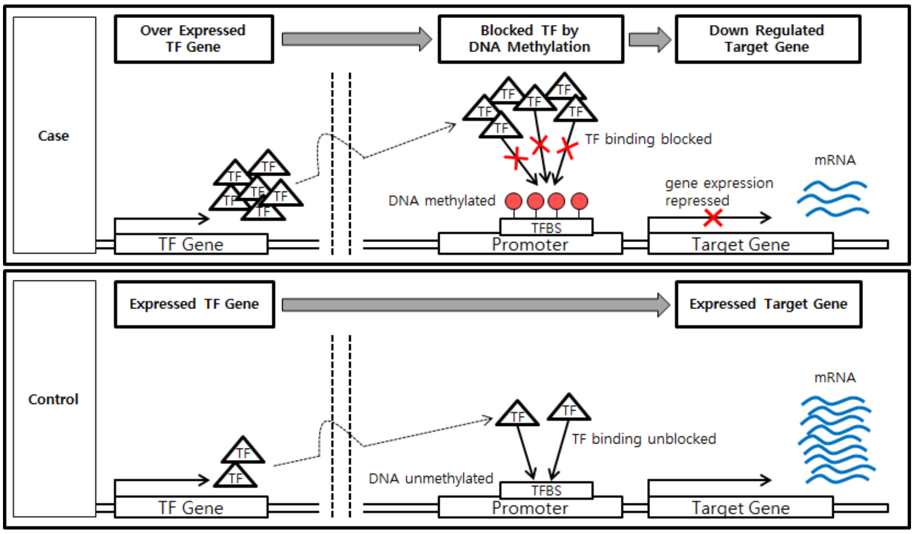


Figure 4.11: Schematic overview of the phenotype-comparative analysis for interference of TF binding by DNA methylation resulting in the suppression of downstream gene expression

overview of the protocol is illustrated in Figure 4.11.

We first identified differentially methylated genes among the downregulated genes, 60 genes in BaB subtype and 52 genes in Lu subtype. Based on the results of the one side standard t-test with a criterion for being significant as p-value<0.005, we observed eight genes with significant hypermethylation in at least one 100bp-bin as follow: *CDH1*, *CLDN4*, *ESRP1*, *GRHL2*, *KRT19*, *PRR15L*, *AKR1B1*, and *PLOD2*. Figure 4.12 shows the promotor regions of the eight genes that are differentially methylated according to the p-values.

Next, for the hypermethylated regions of the eight genes, we generated cell line-specific consensus sequences by assembling short reads mapped to the regions and

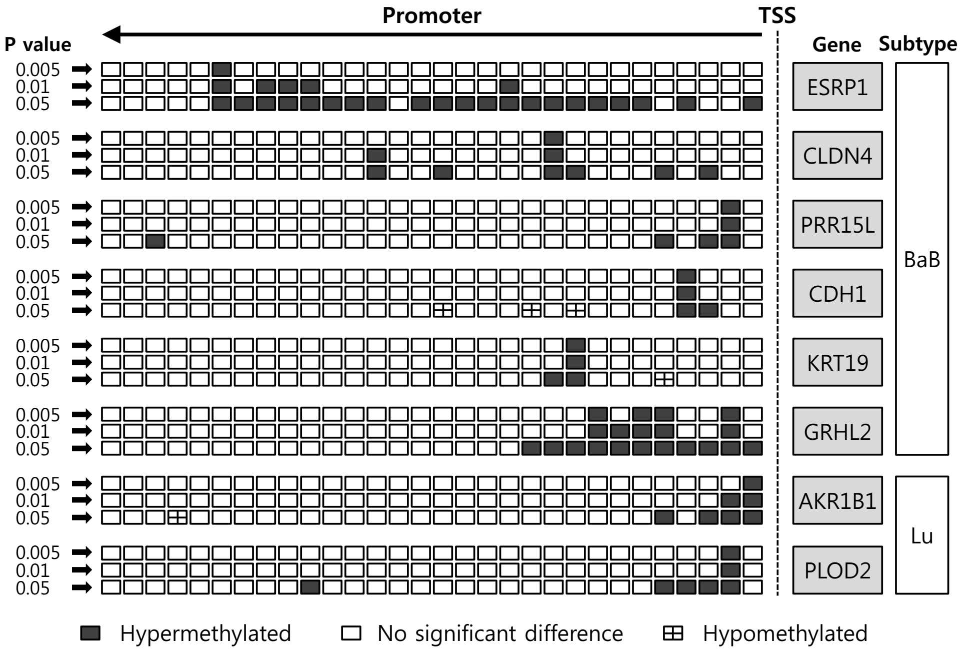


Figure 4.12: Differentially methylated promoter regions of down-regulated genes. Each rectangle in the upstream region means a 100bp-bin.

searched candidate TFs which can be bound to these consensus sequences by match tool (Kel et al., 2003) on the consensus sequences. To exclude the possibility that higher expression of an activator gene might result in upregulation of target genes, we discarded TFs whose expression levels were significantly different across cell lines of different phenotypes.

Table 4.3 summarizes the final selection of TFs and their target genes. TFs appeared in at least 50% of cell lines of the same phenotype (*TFBS Support Rate* in the table is percentage of the number of TF-containing cell lines). Interestingly the genes *CDH1*, *ESRP1* and *GRHL2* have been shown to play critical roles in epithelial-mesenchymal transition (EMT), a process associated with metastatic events in cancer and also highly relevant to tumor progression (Thiery, 2002; Thiery et al., 2009). Lombaerts *et. al.* (Lombaerts et al., 2006) reported that *CDH1* is downregulated by promoter methylation and related to EMT in breast cancer cell lines. A study by Dumont *et. al.* (Dumont et al., 2008) showed that the induction of EMT was accompanied by repression of *CDH1* expression and subsequent DNA hypermethylation at its promoter in basal-like breast cancer. Additionally, recent studies showed that *GRHL2* and *CDH1* in human breast cancer cells were highly correlated and suppressed EMT by repressing expression of the *ZEB1* gene (Xiang et al., 2012; Cieply et al., 2012). *ESRP1* was shown to regulate a switch in CD44 alternative splicing, an event required for EMT and breast cancer progression (Brown et al., 2011). Moreover, there might be potential interplay between target genes. Over-expression of *GRHL2* up-regulated *ESRP1* expression (Xiang et al., 2012), and *GRHL2* was shown to be essential for adequate expression of the *CDH1* and *CLDN4* (Werth et al., 2010). Thus, our approach may be useful to elucidate cell-specific regulatory mechanism using the genome-wide methylation data from the MBDCap-seq.

## 4.4 Discussion

Recent developments in sequencing technologies have made it possible to analyze genome-wide DNA methylation profiles at high resolution. Although altered DNA methylation patterns are a hallmark of cancer, and promoter CGI hypermethylation is known to repress gene expression, only a few studies have examined DNA methylation-gene expression relationships using genome-wide integrated analyses (Ruike et al., 2010; Fang et al., 2011; Sun et al., 2011). Several researchers have attempted to investigate the association of the DNA methylation with the molec-

Table 4.3: Downregulated target gene with transcription factor binding sites on hypermethylated region

| Target Gene | Binding TF | TFBS Support Rate |
|---|---|---|
| CDH1 | SMAD1 | 100.0 |
| CDH1 | FOXO1 | 100.0 |
| CLDN4 | CEBPA | 62.5 |
| CLDN4 | CEBPB | 62.5 |
| CLDN4 | CEBPD | 62.5 |
| CLDN4 | CEBPE | 62.5 |
| CLDN4 | CEBPG | 62.5 |
| ESRP1 | CUX1 | 90.0 |
| GRHL2 | PDX1 | 100.0 |
| KRT19 | PAX6 | 60.0 |
| PRR15L | IKZF1 | 50.0 |
| AKR1B1 | E2F1 | 91.7 |
| PLOD2 | PAX3 | 100.0 |

ular subtypes in breast cancer cells (Bloushtain-Qimron et al., 2008; Holm et al., 2010). However high resolution sequencing data were not used in those studies. To better understand the relationship between DNA methylation and gene expression in breast cancer molecular subtypes, we used next generation DNA methylation sequencing data and gene expression profiles for 30 ICBP cell lines representing molecular subtypes of the disease to perform a systematic analysis.

We first compared genome-wide methylation profiles of breast cancer phenotypes. Although overall DNA methylation profiles were similar in Lu, BaA and BaB, specific genomic regions were differentially methylated among the three subtypes. We then explored computational methods for integrating DNA methylation and gene expression data and started with differentially expressed genes for discovering genes whose expressions were influenced by DNA methylation.

DNA methylation of different genomic regions has recently been associated with altered expression of downstream genes. To better understand possible transcriptional regulatory roles of DNA methylation, we performed a comprehensive study considering distinct genomic regions: CGIs, CGI shores, promoter regions, 1st exons, 1st introns, and 2nd exons. Based on Pearson's correlation coefficients, we verified that the DNA methylation of several genomic regions including CGI and CGI shores were negatively correlated with downstream gene expression.

To investigate combinational effects of DNA methylation in these regions and to identify subtype-specific events, we applied a decision tree algorithm using genes downregulated in each subtype. Interestingly, we found potential combinatorial effects of the first exon methylation and promoter region methylation on the downstream gene expression (BaB subtype) and potential combinatorial effects of CGI methylation and CGI shore methylation (Lu subtype). As gene expression is regulated by many factors, it is difficult to predict gene expression levels using only the DNA methylation profiles. However, the classification accuracy was significantly high enough to elucidate the contribution of each genomic region and combinatorial effects of the regions. We showed that DNA methylation had combinatorial roles on gene expression and the effects of DNA methylation in each genomic region differed among the subtypes. Moreover, our studies further imply that the aberrant DNA methylation state of the TF-associated regions could be another contributing factor to gene repression, a subject of future experimental validation.

It is now well established that different gene expression patterns contribute to breast cancer heterogeneity (Koboldt et al., 2012). In the current study, our integrated analysis further demonstrates that methylation status of different genomic regions may play a key role in establishing transcriptional patterns in three molecular subtypes of human breast cancer. Understanding the functional impact of distinct regions of DNA methylation on gene expression patterns may provide additional insight into breast cancer progression and response to therapy, both critical for improving management of the disease.

# Chapter 5

# Detecting multiple SNP interaction via evolutionary learning

## 5.1 Background

Genome-wide association study (GWAS) examines genetic variations on the whole genome of individuals and investigates how the variants frequently occur in population with a particular phenotype such as disease. The main purpose of the GWAS is to identify the genetic variations which influence to phenotypic changes or are susceptible to diseases. One of the most popular variants to use in the GWAS is single-nucleotide polymorphism (SNP). SNPs were relatively easy to be identified, and many people believed that the cause of disease would be discovered by the variants. In reality, there have been a lot of research to capture the genetic variants which are statistically associated disease or traits, and as a result of GWASs, it has been reported that hundreds of loci are associated for more than 70 common diseases and traits (Donnelly, 2008).

However, comprehensive understanding for the relationship of genotypes to phenotypes, is still challenging. The complex traits including cancers and diabetes are believed to be affected by the interactions of multiple genetic factors (Cordell, 2009). In many cases, the single genetic variants did not fully explain a cause of the complex disease.

To understand the complexity of mapping from genotype to phenotype, many researchers have focused on genetic interaction and relationships of more amount of variants, instead of a single genetic marker (Heidema et al., 2007; Cordell, 2009). Especially, machine learning approaches could be a useful solution of the problem (Szymczak et al., 2009; Moore et al., 2010). For example, logic regression and decision trees could be applied for the analysis of the interaction of variants (Ruczinski et al., 2004; Fiaschi et al., 2010). Another widely used technique is MDR (multifactor dimensionality reduction) approach (Ritchie et al., 2001) which has been developed with the idea of CPM (combinatorial partitioning method). However, these have limitations in efficiently handling higher order interactions from a large number of SNPs.

Here, we address the multiple SNP associations to disease, by the construction of a classifier based on evolutionary learning. One of the important steps to improve the performance of a classifier is to identify the informative feature sets. Especially, in the association study, the number of features is very high, and in the case of concerning all of the multiple combinations of the attributes, most of computational learning algorithms might fail to efficiently control the large-scale datasets. We introduce a concept of evolutionary learning to identify higher-order combinatorial features which are relevant to the class discrimination, from the combinatorial search space. Generally, evolutionary learning well-approximates solution to complex problems which are difficult to optimize mathematically. For the genetic association studies, several research has been accomplished by the evolutionary learning, and showed that it could be applied successfully (Namkung et al., 2007; Moore and White, 2007; Nunkesser et al., 2007; Clark et al., 2008; Yang et al., 2010).

We propose a method to find association of multiple SNPs and a disease, and to predict a disease by the variant information. Firstly, we applied the approach to a simulation data and verified the approach could be useful to find the SNP interactions. After that, we identified the combinatorial effects of multiple SNPs on T2D in Korean population. In our evolutionary algorithm, a single individual is encoded by the form of explicit rules which are formulated for certain values of the attributes, and the whole population evolves to the final rule-set with a good fitness. In the learning process, the evolutionary computation can solve the problem efficiently by avoiding exploring the whole search space and leading to identify higher-order SNPs with strong association to a phenotype. The resulting rule set is able to correctly recognize instances and discriminate them to target concepts well. As a result, the model can classify the instances by combination of the survived rules after evolutionary learning, and the rules can be considered as informative multiple factor interactions associated to a disease.

## 5.2 Materials and methods

### 5.2.1 Identifying higher-order interaction of SNPs

The evolutionary computation approach, particularly learning classifier system (LCS) has successfully applied to induce a set of classification rules in a given environment (Bernado-Mansilla and Garrell, 2003; Sigaud and Wilson, 2007; Fernandez et al., 2010). The LCS searches the space of possible rules, guiding the search for better rules by evolutionary computing techniques. Our main goal is similar to the technique. We construct an evolutionary learning method guided by a gradient descendent algorithm, to induce a set of classification rules from SNP data with complex traits. The detail explanation follows.

### Structure of the individuals

Suppose that $X = \{X_1, X_2, ..., X_n\}$ is a dataset of $n$ samples, and each sample $X_i$ is composed by $k$ features, that is, SNP loci, and class value $y_i \in \{normal, disease\}$. The input value of each feature in the SNP data can take one of the following three states: (1) homozygous major form, (2) heterozygous, and (3) homozygous minor form. The structure of the individuals are expressed as a combination of SNP information. For example, an individual is represented from the conjunctive form of the multiple SNP association as follows:

$(SNP_1 = 3) \bigwedge (SNP_2 = 2) \bigwedge (SNP_3 = 2)$

It means that the $SNP_1$ is hetero, $SNP_2$ is homo minor form and $SNP_3$ is also homo minor form.

## 5.2.2 Algorithm Description

The algorithms steps are summarized in Table 5.1 and Figure 5.1. More detail is given on individual steps in following subsections.

### Initialization

In the evolutionary learning, population is defined by a set representing higher-order interaction among SNPs. The initial population is consisted by individuals randomly generated with chromosome length $l$. The population size $s$ is decided empirically and the initial weight $w_j$ of the individual $j$ $(0 < j < s)$ is randomly assigned with

Table 5.1: Overall learning procedure

**Main Learning Procedure:**

1. Randomly generate a population and initialize $s$ individuals with weights $w$s. The length of chromosome $l$ is user-specified. The weight (fitness) $\mathbf{w}$ is randomly initialized with a small value.

2. Train the weight value of each individual iteratively using instances. The weight values are updated and assigned by a gradient-descent algorithm. The learning procedure in step 2 is terminated when the weights are converged after repetition of a number of epoch.

3. The evolutionary process begins. Remove individuals with worst fitness from population. The individual is worse as its fitness is closer to zero. Theses are replaced by newly generated individuals. The offsprings are reproduced by one of four ways in user-specified proportion.

   (a) Inherit $r$ individuals whose $e_j$ is -1. (elitism)

   (b) $\alpha$ individuals should be generated by the crossover operator. By selection strategy (ranking selection), select two individuals and crossover them.

   (c) Mutate $\beta$ individuals in the parents.

   (d) Randomly generate $s - r - \alpha - \beta$ individuals

4. Go to Step 2 until convergence after the number of generation.

$r$ is a parameter for the number of removing individuals

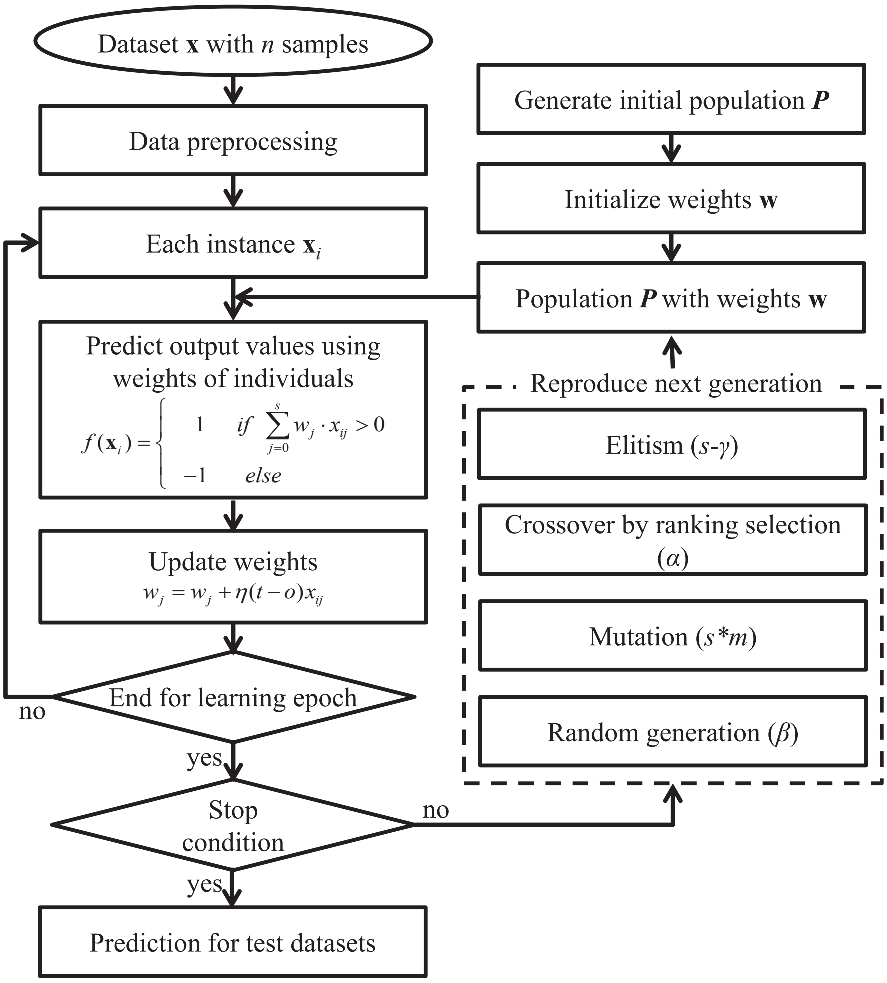

Dataset **x** with *n* samples
Data preprocessing
Each instance $\mathbf{x}_i$
Predict output values using weights of individuals
$f(\mathbf{x}_i) = \begin{cases} 1 & if \sum_{j=0}^{s} w_j \cdot x_{ij} > 0 \\ -1 & else \end{cases}$
Update weights
$w_j = w_j + \eta(t-o)x_{ij}$
End for learning epoch
no
yes
Stop condition
no
yes
Prediction for test datasets
Generate initial population ***P***
Initialize weights **w**
Population ***P*** with weights **w**
Reproduce next generation
Elitism ($s$-$\gamma$)
Crossover by ranking selection ($\alpha$)
Mutation ($s$*$m$)
Random generation ($\beta$)

Figure 5.1: Flow chart for our evolutionary learning method. The most fitable individual is searched by the iterative learning.

a small value (-1 $< w_j <$ 1).

### Weight Update and Evaluation

Each individual has a weight value which means how informative the chromosome is to classify the samples. That is, the weights for individuals are considered as their fitness and the bigger weight on an individual mean mores informative to classify the instances. To determine and update the fitness for each individual, we introduce a gradient descendant rule as follows:

$$w_j = w_j + \eta(t_i - f(\mathbf{x}_i))m_{ij}, \tag{5.1}$$

where $w_j$ is a weight value for $j$-th individual and $t_i$ is a target class in the $i$-th training instance. $m_{ij}$ is a variable whether the all values of attributes within the $j$-the individual is matched to those in the $i$-th instance.

$$m_{ij} = \begin{cases} 1, & \text{if all values are identical} \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

$f(\mathbf{x}_i)$ is a predicted output value of the $i$-th training instance by our model and determined as follows:

$$f(\mathbf{x}_i) = \begin{cases} 1, & \text{if } \sum_{j=0}^{s} w_j \cdot m_{ij} > 0 \\ -1, & \text{else} \end{cases} \tag{5.3}$$

The difference between the predictions and the target values specified in the training sequence is used to represent the error of the current weight vector. The target function is optimized to minimize the classification error. The weight values are evaluated against a sequence of training samples and are updated to improve the classification accuracy. The weight update processes are repeated until it would be converged after the number of epoch.

Using the learning scheme, we find most informative individuals for classification, that is, the absolute value of their weights is large.

### Removing and Reproduction

During each successive generation, a proportion of the existing population is selected to be survived in the next generation. We adopted individual replacement strategy in each generation during the evolution processes. Basically, the highly weighted individuals should be selected and the others dismissed. It is a similar concept with elitism in a conventional genetic algorithm. We measure the fitness of each individual and preferentially select $r$ best solutions. The $r$, the number of individuals to be survived, is determined by a threshold $\theta$.

$$e_j = e(w_j) = \begin{cases} -1, & \text{if } |w_j| < \theta \\ 1, & \text{otherwise} \end{cases} \tag{5.4}$$

, where $|\cdot|$ means a absolute value. Then, individuals whose $e_j$s are 1, is survived and the $s$ - $r$ individuals are removed. After that we generate new individuals as much as removed in the step.

$s$ - $r$ individuals are reproduced by three ways in the next generation. The first is random generation. As similar to the the process for making initial population, we can construct new individuals randomly. Another ways are bring from conventional genetic operators, crossover and mutation. We select two individuals by ranking selection and can recombinate them in a random position. $\alpha$ and $\beta$, the number of individuals to be generated by crossover and mutation, respectively, are determined as follows:

$$\alpha = \lambda(s - r) \tag{5.5}$$

$$\beta = s - r - \lambda(s - r) \tag{5.6}$$

, where $\lambda$ is a crossover rate. For the mutation operator, there exists two kind of alteration. We choose $\kappa$ individuals and substitute a gene to another gene. There is the other base mutation rate $\rho$. It change a value of a variable in a selected individual for mutation to other one, so it help to be exploited in the search space by giving a variation in the combinatorial factors of the individual.

$$\beta = \mu + \kappa = \rho\beta + (1 - \rho)\beta \tag{5.7}$$

, where $\rho$ is the base mutation rate, and $\mu$, the number of individuals to be occurred base mutation, is determined by the $\rho$ value.

#### Iterative Learning Procedure

To select interesting rules from population, that is, the sets of the possible rules, we iteratively reproduce the individuals in progress on generation. The individuals are required to satisfy the specified fitness function and are survived only if they are adapted in the environments, that means they are how much informative to classify the training dataset.

By repeating the procedures until convergence (Table 5.1), the model can classify the normal and disease samples well, and identify higher-order interactions of SNPs.

### 5.2.3 Dataset

Genotyping and clinical information of Korean individuals was produced as a part of Korean Association Resource (KARE) project by Korean Centers for Disease Control and Prevention (Cho et al., 2009; Hong et al., 2012). The cohort study was examined for 8842 individuals at Ansan and Ansung area, aged 39 to 70. The genotyping was conducted using Affymetrix Genome-Wide Human SNP array 5.0. In the clinical information, we investigated the concentration of glucose, diagnosis, and drug treatments. If a person have been an experience to take a diagnosis for the diabetes, we decided the person have a diabetes. Also if plasma glucose is same to or over 126 (mg/dl) in no caloric intake or two-hour plasma glucose is same to or over 200 (mg/dl), then we considered the person a diabetes case. Conversely, the criterion for the normal controls are the plasma glucose with no calory intake is under 100, two-hour is under 140, and no experience for diabetes diagnosis or insulin intakes.

### Odds ratio

The odds ratio is used to measure a relative risk in a specific genotype comparing to another one. It was calculated as follows:

$$oddsratio = \frac{p_1(1-p_1)}{p_2(1-p_2)} \tag{5.8}$$

, where $p_1$ and $p_2$ are probabilities that an individual having the selected SNP rules exists in the disease group and normal group, respectively. If an odds ratio is greater than 1, the events is more likely to occur disease. That is, the odds ratio which is significantly higher than 1, means the higher-order SNPs are associated with disease.

The p-value is measured by random combination of SNPs. We generated 100,000 SNP rules randomly, and calculated odds ratio in each rule. Then we checked the probability that the odds ratio for the selected rules occurs by chance.

## 5.3 Results

### 5.3.1 Identifying interaction between features in simulation data

To verify our approach can find the interaction of features, we tested the method using simulation data. Suppose that the simulation data $X_i = (x_1, x_2, ..., x_{10}, class)(1 \leq i \leq 1,000)$ is composed of 10 attributes, $x_j \in 0,1(1 \leq j \leq 10)$. By gibbs sampling, we generated the data with following conditions:

$$\begin{aligned}
&P(x_1 = 0) = 0.6\\
&P(x_2 = 0) = 0.6\\
&P(class = 1|x_1 = 1 \wedge x_2 = 1) = 0.8\\
&P(class = 1|x_1 = 1 \wedge x_2 = 0) = 0.3\\
&P(class = 1|x_1 = 0 \wedge x_2 = 1) = 0.3\\
&P(class = 1|x_1 = 0 \wedge x_2 = 0) = 0.2
\end{aligned} \tag{5.9}$$

$x_3$ and $x_4$ have same probabilities with $x_1$ and $x_2$, respectively and the others are randomly generated (uniformly distributed). Table 5.2 shows the finally selected interactions by our approach using the simulated data. As we expected, our method can find the informative interactions of features for the classification with around 0.70 classification performance by 10-fold cross validation. The set of $x_1$ and $x_2$ were selected as the most highly ranked interaction. Also, the pairs of $x_3$ and $x_4$ similarly had big weight values after the learning by our approach. Although a state of the art classifier, SVM, has a little higher accuracy (0.734), the algorithm does not provide which features are important for the classification. Moreover, it is impossible to detect the combinatorial effects among the genetic variants, but our method can it.

Table 5.2: Identified interaction in simulated data

| Interaction | Weight |
|---|---|
| $x_1$=1, $x_2$=1 | 0.91 |
| $x_1$=1, $x_2$=0 | -0.89 |
| $x_3$=1, $x_4$=1 | 0.81 |
| $x_3$=0, $x_4$=1 | -0.79 |
| $x_3$=1, $x_6$=1 | -0.79 |

### 5.3.2 Identifying higher-order SNP interactions in Korean population

Korean population might be specific associative characteristics to a disease. Since we confirmed that our proposed method would be adequate to be find combinatorial effects of SNPs in genome-wide association study, at last we searched the multiple SNP interaction in Korean population using our method.

For the preprocessing, we firstly carried out hardy-weinberg test (HWE), then filtered out uninformative SNPs (p value $< 0.000001$). Then we removed SNP attributes where minor allele frequency (MAF) is less than 0.01. Then for each SNP, the p-value was calculated based on a chi-square test. We also filtered out significant SNPs (p value $< 0.05$). After the preprocessing, the number of attributes was decreased to 6459.

The main purpose of our approach is to identify higher-order interaction of multiple SNPs, but it can be run as a classifier. Also, it is required to check the classification performance for selecting highly discriminative combination of SNPs. Table 5.3 shows classification accuracy in our method. Using 10-fold cross-validation, the classification accuracy was around 90% when we evaluated the performance along the chromosome length. We also carried out other classification algorithms using the same datasets and compared the accuracy (Table 5.3). Even though it had a little difference with the interaction length to be examined, we obtained better or competitive performance to the results of other general classifiers. Usually, tree-based classifiers can be used to know which factors affects to the classification. However, in the dataset, the tree-based methods were shown much lower classification accuracy, 61.11% with decision tree (ID3) and 70.66% with random forest, which is considered as a combination of decision rules in classification tree forms. The classification accuracies of other approaches were also about 73.59% with an instance based classifier (k-nearest neighbor, kNN) and 71.38% with logistic regression. Only RBF network and SVM achieved the similar accuracy to our method. However, these two algorithms do not provide which factors significantly affect to the classification. The results mean that, our approach can find higher-order interactions of SNPs by choosing the highly-weighted individuals from the learned models, along the chromosome lengths.

Table 5.3: Classification performance in KARE dataset

| Order ($l$) | Accuracy |
|---|---|
| $l$=2 | 91.20 |
| $l$=3 | 91.40 |
| $l$=4 | 89.16 |
| Decision Tree (ID3) | 61.11 |
| Decision Tree (C4.5) | 60.22 |
| Random Forest | 70.66 |
| kNN (k=10) | 73.59 |
| SVM | 94.81 |
| RBF Network | 92.83 |
| Simple Perceptron | 67.11 |
| Logistic Regression | 71.38 |

In each experiment, we selected top SNP combinations from the ranking of their weights, and subsequently, we evaluated significance of the interactions through the odds ratio and the chi-square test. Table 5.4 shows top 10 interactions as an experimental result with order 3, and Figure 5.2 show the interaction map. The highly positive value of the weight implies the interaction can be a big effect to T2D, and negative means it is affectable to be classified to the normal sample. The table presents that the positively weighted interactions all have the high (>1) odds ratio. Conversely, the interactions with negative weight values low (<1) odds ratios. That is, the results suggest that the positively-weighted interaction is able to be a candidate for the T2D risk factors. In addition, the interactions were significantly distinguishable between case and control data by a chi-square test. the p-values by a chi-square test were significantly low in the whole selected interactions.

Table 5.4: Highly ranked SNP interaction

| SNP interaction | | | weight | odd ratio | p-value (chi square test) |
|---|---|---|---|---|---|
| SNP_A-4196226 TT | SNP_A-2038226 CC | SNP_A-1861290 GG | -3.7 | 0.754 | 3.09e-4 |
| SNP_A-1963560 CC | SNP_A-4222651 TT | SNP_A-2032424 GG | -2.78 | 0.738 | 5.77e-5 |
| SNP_A-2144088 GG | SNP_A-2055282 CC | SNP_A-2293836 AA | 2.52 | 1.416 | 6.49e-6 |
| SNP_A-2182681 GG | SNP_A-4223259 TT | SNP_A-2033011 GG | 2.18 | 1.366 | 7.42e-5 |
| SNP_A-2269625 CC | SNP_A-2257007 CC | SNP_A-1788186 AA | 1.86 | 1.352 | 1.91e-4 |
| SNP_A-2178766 CC | SNP_A-1932380 CC | SNP_A-2224407 GG | -1.84 | 0.707 | 7.93e-6 |
| SNP_A-1842269 TT | SNP_A-2205381 CC | SNP_A-1982225 AA | 1.84 | 1.209 | 1.31e-2 |
| SNP_A-1829387 TT | SNP_A-1861385 CC | SNP_A-4283627 AA | 1.74 | 1.387 | 2.86e-5 |
| SNP_A-1802450 CC | SNP_A-2048106 GG | SNP_A-2033011 GG | 1.64 | 1.387 | 1.86e-5 |
| SNP_A-1869508 TT | SNP_A-1844690 AA | SNP_A-1884338 CC | 1.64 | 1.348 | 3.00e-4 |

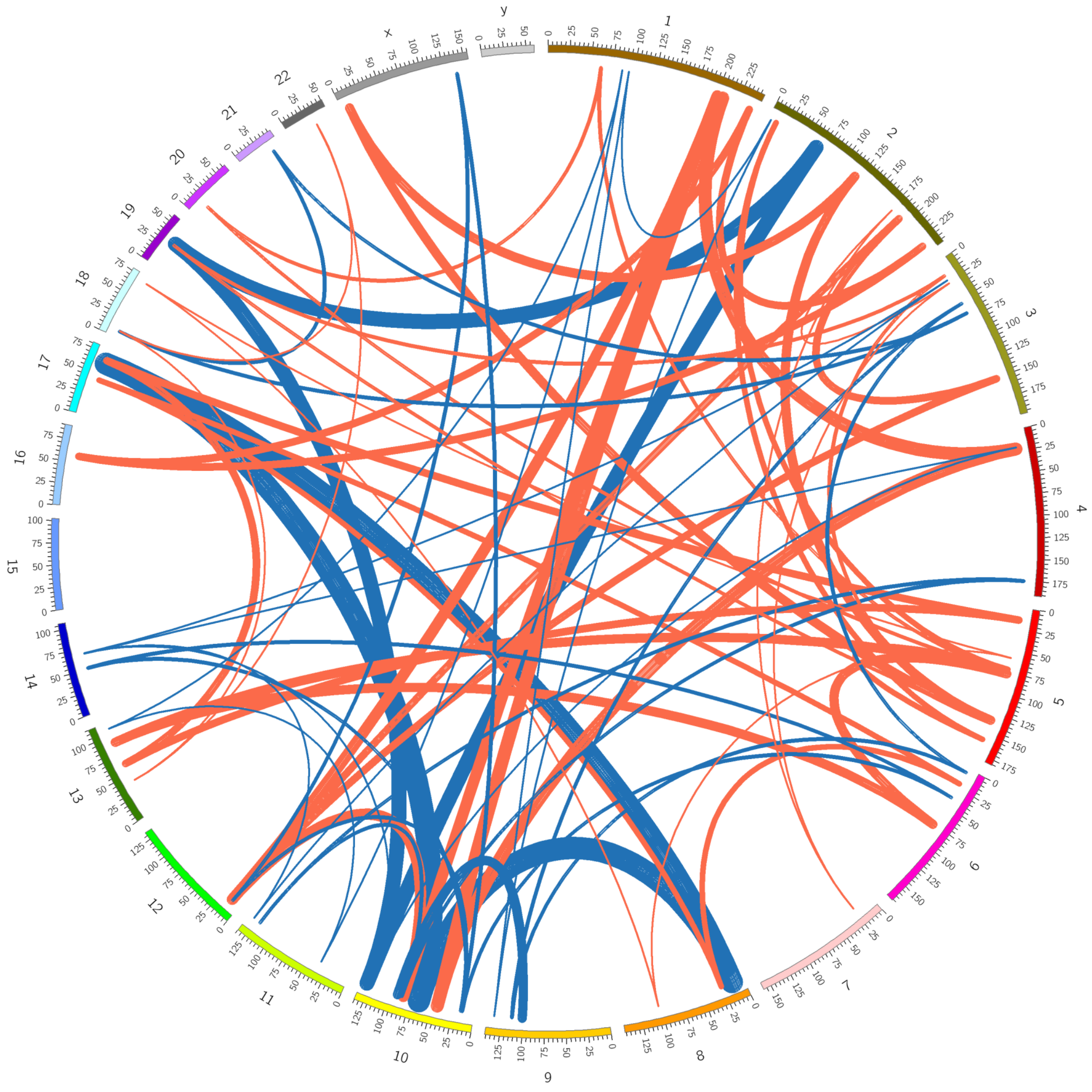


Figure 5.2: SNP interaction map order 3. The thickness of the lines means weights of the interactions. Blue and red colors mean negative and positive weights, respectively.

Interestingly, our results showed that the sequence variation could have much clear association with the higher-order interaction, even although it did not show the strong evidence in single-SNP analysis. Figure 5.3 shows the results for the top 5 ranked interactions. The p-values of the identified interactions were clearly lower than those in single variants within the interaction by our experiments. For an instance, the firstly ranked interaction, SNP_A-4196226, SNP_A-2038226 and SNP_A-1861290 did not show clear association with diabetes as a single variant. The p-values for the single SNP were 0.06, 0.04, 0.02, respectively. However the combination of these was definitely stronger effects to a disease with 3.09e-04 p-value.

For further validation, we randomly generated 1,000 interactions which consist of 3 SNPs and choose the interactions whose number of the matched to instances are more than 10. Then we measured their p-values by the chi-square test. Figure 5.4 shows the p-value comparison between top 100 interactions in our results and the randomly generated set. It shows that the interactions in our results are much more significant. When we carried out a t-test to clarify how these two sets are different, the p-value by the t-test was 9.79e-133.

## 5.4 Discussion

We presented a method to identify higher-order interaction of multiple variables. The study to identifying the higher-order interaction of genetic variants is necessary to find the multiple causal factors, contribute to complex diseases. Although the analysis of multiple factor interaction should be important in understanding complex traits, however, it is computationally infeasible to combinatorially explore all high-order interactions among the SNPs in a genome-wide association study. Previously several studies reported on findings of interactions among genes to be

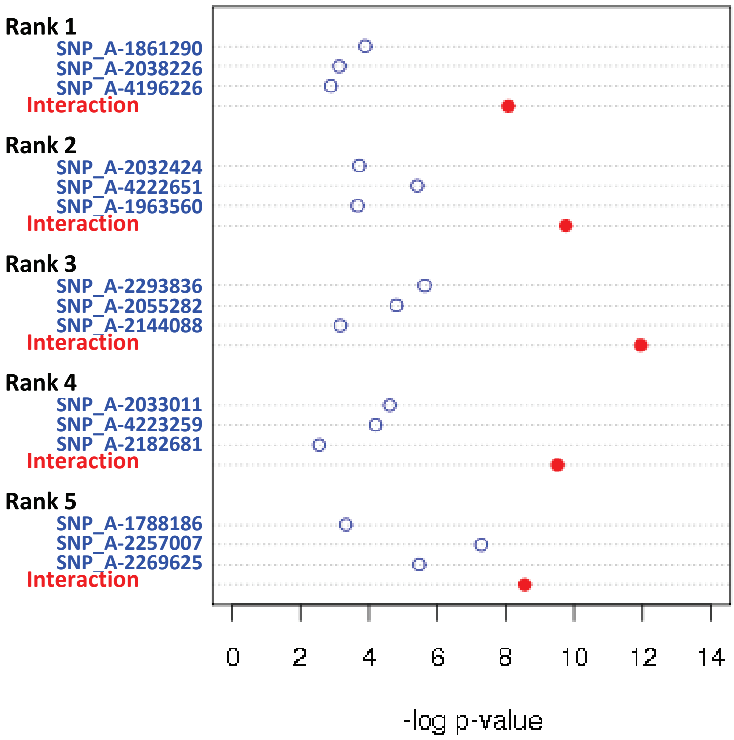


Figure 5.3: Dotchart for comparison between single variants and their interaction. Empty circle is for a single variant and filled is for the interaction.

important contributors to certain phenotypic variation. However, in addition to the variants of genes which directly changes protein function, the genetic alteration may be located in genomic or epigenetic regulatory regions. These can also affect to the gene regulation and abnormality in cellular processes.

We used evolutionary learning to search the combinatorial feature spaces. Generally evolutionary computation finds a good solution by a guidance from a fitness and genetic operators. Using the concepts, we could find a solution, coherent group

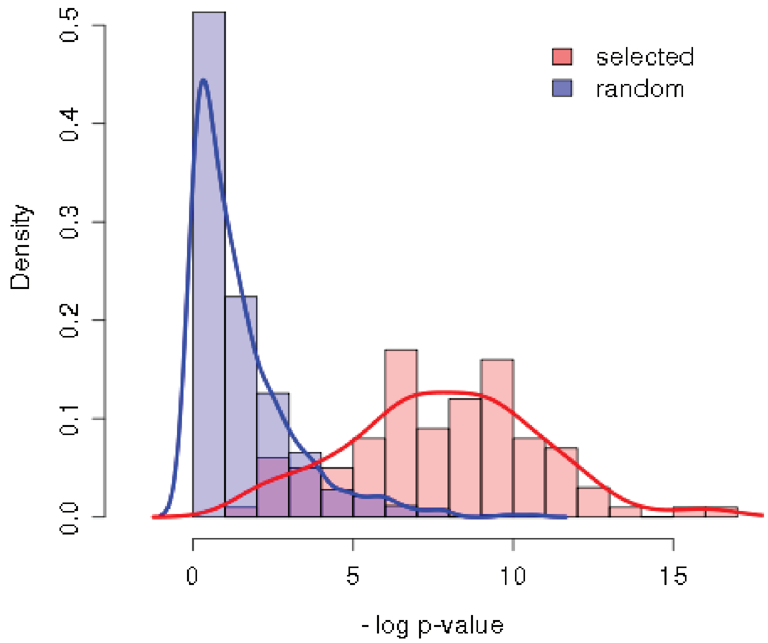


Figure 5.4: Comparison between the interaction by our approach and random selection. Red is a histogram for the identified interaction by our method, and blue is random selection.

of interrelated variants, associated to a disease effectively. When we examined every possible case, the search space is too big. For example, If the number of attributes is 6459 and the combinatorial order is increased from 2 to 5, the number of possible combinatorial cases are 2.09E07, 4.49E10, 7.25E13, and 9.35E16, respectively. However, we searched only cases less than 1.00E6 in every experiment and could find reasonable high order interaction associated to disease. Our genetic association studies for complex traits can be applied to a systems genetics studies integrated with other information, such as environmental factors, copy number variation, clinical information, and so on. The systems genetics approach helps to yield a detailed map of genetic and other variants, including environments, associated with phenotypes. Our proposed methods can easily add these factors in steps to generate individuals, and find their effects to a disease. Also, the evolutionary learning in our approach make it possible to control the large datasets with a explorative search space. so a number of factors can be supplied in the consistent algorithm.

In our experiments, we did not reflect biological knowledge or genetic relationships. Depending on a experimental purpose, these information can be reflected in the process on generation of individuals or in the fitness function. Or it is also possible to construct a model with genetic relationships by measuring linkage blocks or conducting a transmission disequilibrium test from datasets.

In addition, the analysis of the interaction accompanies several issues including information loss with missing values. But our approach does not require imputation of the missing values, and it can be run by denoting these missing values as don't care symbols or mismatched symbol.

Sometimes, a sampling approach is an efficient method to find an optimal solution in a large datasets. However the datasets would be too sparse, especially in case of higher-order combination of variants. So we should randomly generate some of the individuals, instead of sampling from training datasets. In addition, if we want to search the interactions between just two variables, it might be not necessary to use crossover or mutation. It could be possible to find the fittest one, by random sampling in the reproduction processes. But in the interactions of multiple variables, it would be efficient to use these operators.

In each experiment, the chromosome length was constant. If the experiments are carried out to identify interactions with a variety length at one time, the individuals with a small length are more likely to be matched to a datasets, so it can be much bigger weight values. However, our approaches can be easily expanded to a method for identifying the interactions of variable length. One way is to normalize the fitness value by the chromosome length. Then we can find the interactions of various orders, resulted from individuals with diverse lengthes. Another way is to learn from lower to higher order by turns, and then to re-learn and classify based on the finally survived individuals in each step.

By the characteristics of evolutionary learning, our results would not be global optima. But it is definitely valuable. Our purpose is not to find one optimal coherent variant set associated to a disease. Also it might be impossible to be expressively provided that the complex traits are caused only by a little number of factors. The reason of the disease occurrence is not simple. Therefore, we detect interactions which may be local optima and provide the candidates to help to find sets of the risk factors.

Recent advances in high-throughput sequencing provide a variety of datasets. The sequencing datasets may shed light for a new finding in the GWAS, and whole-genome or whole exome sequencing has been used to search the genetic cause of diseases. Despite of the considerable progress in the sequencing technologies and their analysis strategies, the common variations identified by GWAS account for only a small fraction of disease heritability and are unlikely to explain the majority of phenotypic variations of common diseases. Our approach can be usefully applicable to the sequencing datasets. The sequencing technologies have detected millions of novel variants. Although big size of dataset by lots of reads and variants is another challengeable problem, our approach can be a method to solve the problem by effectively searching the combinatorial feature space based on the evolutionary learning. It can be a effective method to systematically control exploration of a lot of variants provided by next generation sequencing technologies for GWAS. Also, the sequence datasets have a large proportion of missing data, but our method can be resistable.

Our approach suggests the analysis of GWAS datasets offers a useful strategy for identifying causal genes and potential candidates in human diseases. Study for interaction of the genes or genomic regions would help to elucidate mechanism of the complex traits and to control and treat disease. Some of our results do not show clear relationships and some of these may be still biologically questionable, why the combination is highly weighted and how there play a role in disease. For the much clear understanding, relevant functional studies should be carried out. Moreover, by applying phased haplotype information, we will detect much relevant sets for variants (Tewhey et al., 2011).

# Chapter 6

# Identifying DNA methylation modules by probabilistic evolutionary learning

## 6.1 Background

Genomics mainly aims to find genetically associated markers with a phenotype. Based on DNA sequences, the researchers search causal effects to biological processes including gene regulatory mechanism and disease. Although several risk factors were identified by the association studies, the genetic variants do not fully explain the abnormal regulation, since the biological regulatory mechanism can be affected by many other factors, as well as DNA sequence modification (Jones and Baylin, 2007; Sadikovic et al., 2008; Handel et al., 2010; Sandoval and Esteller, 2012).

Epigenomics refers to a study for regulation of various genomic functions that are controlled by another partially stable modification, not DNA sequence variants (Bonetta, 2008). Among these, DNA methylation, which typically occurs at CpG dinucleotide by DNA methyltransferase (DNMT) enzyme, is a crucial epigenetic regulatory mechanism in cellular processes. The DNA methylation of CpG site mostly cause silence of the downstream gene, so the enrichment of the differentially methylated DNA fractions can contribute to specific abnormalities, including complex diseases (Robertson, 2005; Portela and Esteller, 2010; Jones, 2012). Especially, with an advent of microarray and next generation sequencing (NGS) technology, many researchers have carried out genome-wide DNA methylation profiling studies (Laird, 2010; Hill et al., 2011; Rhee et al., 2013), and the genome-wide studies have reported that lots of genomic regions are differentially methylated in normal and abnormal cells (Cheung et al., 2010; Toperoff et al., 2012; Walker et al., 2011).

However, it is well-known that a complex disease is generally caused by combinatorial dis-regulatory effects of multiple genes (Hirschhorn and Daly, 2005; Janssens and van Duijn, 2008; Kiezun et al., 2012). That is, the errors of biological processes is not caused by alteration of an individual methylation level. Recently, Easwaran et al suggested a concept for DNA hypermethylation modules which preferentially target important developmental regulators in embryonic stem cells (Easwaran et al., 2012). They found the set of genes by the DNA methylation would be contribute to stem-like state of cancer. Horvath et al. studied aging effects of DNA methylation and showed there exist co-methylated modules related to aging in human brain and blood tissue (Horvath et al., 2012).

Here, we identity combinatorial modules of DNA methylation sites associated to human disease by an evolutionary learning approach. The evolutionary algorithms can approximate solution well in lots of problems (Kumar et al., 2010; Deb and Datta, 2010; Joung et al., 2012; Wang et al., 2013). It generates new population through iterative updates and selection by a guided search process in a feature space. We utilized an estimation of distribution algorithm (EDA)-based learning approach for identifying combination of cancer-related DNA methylation sites. In the EDA algorithm, the population is evolved according by probabilistic distribution in the selected individuals without conventional genetic operators such as crossover and mutation. It has been known that EDA efficiently and effectively provide answers in combinatorial optimization problems (Chen et al., 2009; Zhou et al., 2009; Shim et al., 2013; Ceberio et al., 2013). The EDA has been previously applied in several biological research, and it has offered promising results for complex problems, in where other methods fail to find good solution (Pal et al., 2006; Santana et al., 2010; Shelke et al., 2013).

In this study, we investigated DNA methylation modules relevant to cancer, using the DNA methylation profiling datasets produced by microarray- and sequencing-based approaches. The experimental results show that our method can find the DNA methylation modules well related to cancer.

## 6.2 Methods

### 6.2.1 Evolutionary learning procedure to identify a set of DNA methylation sites associated to disease

EDAs evolve a population to find optimal solution probabilistically. The initial population is composed by constructing individuals at random. The individual represents higher order interaction of the methylated sites. The population size $m$ is decided empirically and the initial weight $w_j$ of the individual $j$ $(0 < j < m)$ is randomly assigned with a small value (-1 $< w_j <$ 1).

In the evolutionary process, each individual is evaluated how the interaction is discriminative for the datasets. Then, the better individuals are selected and the dependency tree fitted to the selected individuals, is build. New individuals of the next generation are generated using the probability distribution within the tree structure, and replace the previous individuals. The overall procedure follows:

1. Set $g \leftarrow 0$
2. Initialize population $X(g)$ by random generation
3. Evaluate individuals in $X(g)$
4. Select a set of individuals by tournament selection from $X(g)$
5. Construct a dependency tree $G(g)$ by measuring Kullback-Leibler divergence between variables
6. Parameter learning using probability distribution of the selected set
7. Generate a new individuals by sampling with joint distribution from the $G(g)$, and create new population $X(g+1)$
8. Set $g \leftarrow g+1$
9. If the termination criterion is not met, go to 3

More details for steps 3 and 5 are explained in following sections.

### 6.2.2 Learning dependency graph

The dependency tree is built from the selected individuals by searching conditional dependencies between random variables. Then the model is optimized by a series of incremental updates (Pelikan, 2006; Pelikan et al., 2007). More details follow:

Suppose that $X$ is population and $X = \{X_1, X_2, ..., X_n\}$ is presented as a vector of variables with $n$ features, that is, DNA methylation sites. The probability distribution is represented by a joint probability $P(X_1, X_2, ..., X_n)$ as to:

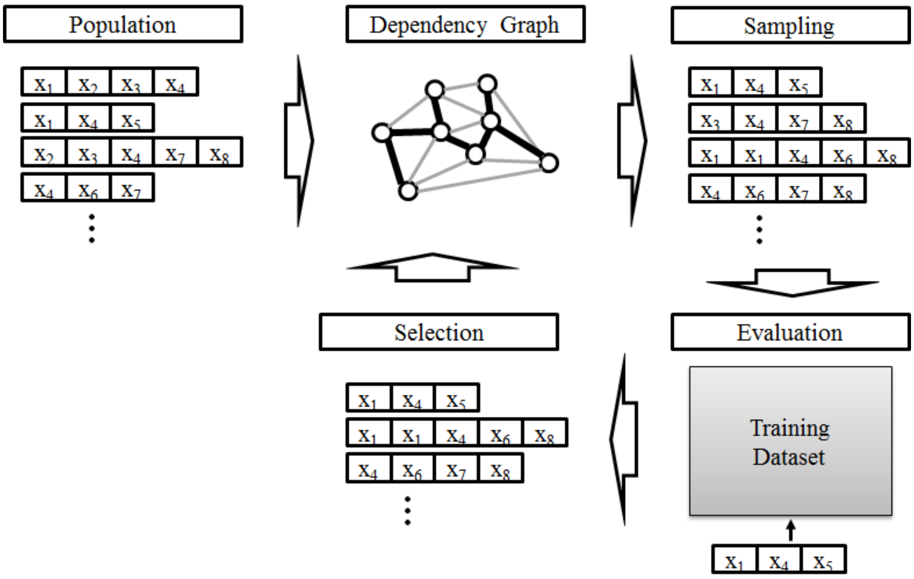


Figure 6.1: Schematic overview for probabilistic evolutionary learning to identify DNA methylation module, Iterative evolutionary learning.

$$\begin{aligned} P(X) &= P(X_1, X_2, ..., X_n) \\ &= P(X_1|X_2, ..., X_n)P(X_2|X_3, ..., X_n)....P(X_{n-1}|X_n)P(X_n). \end{aligned} \tag{6.1}$$

However, it is hard to measure all the joint probabilities exactly when $n$, the number of variables, is large. Thus it needs to approximate the probability distribution. For the purpose, in this study, we used a dependency tree, and the distribution is approximated as follows:

$$P(X_1, X_2, ..., X_n) = P(X_r) \prod_{i \neq r} P(X_i|X_{pa(i)}), \tag{6.2}$$

where $X_1, X_2, ..., X_n$ are random variables, $r$ is an index of root node, and $pa(i)$ denote the index of parent node of $X_i$.The tree structure is built by searching based on Kullback-Leibler divergence between two random variables. The dependency graph is optimally constructed in a direction to maximize total mutual information as follows:

$$argmax_{r,pa} \prod_{i \neq r} I((X_i)pa(i)) \tag{6.3}$$

$$\begin{aligned} &I((X_i)pa(i)) = \\ &\sum_x \sum_y P(X_i = x, X_{pa(i)} = y) log \frac{P(X_i = x, X_{pa(i)} = y)}{P(X_i = x)P(X_{pa(i)} = y)} \end{aligned} \tag{6.4}$$

The complete graph $G$ searches the maximum spanning tree, and then the best dependency tree is constructed.

For parameter learning, the most likely values are calculated from the frequencies in the selected individuals. That is, the model parameters are represented as a marginal probabilities in a root node and conditional probabilities in the other nodes. The marginal probabilities in root nodes and the conditional probabilities in child nodes are calculated as:

$$P(X_r = x) = \frac{m(X_r = x)}{N}, \tag{6.5}$$

$$P(X_i | X_{pa(i)}) = \frac{m(X_r = x)m(X_i = x, X_{pa(i)} = y)}{m(X_{pa(i)})}. \tag{6.6}$$

### 6.2.3 Fitness evaluation in population

Each individual has a fitness value which means how informative the chromosome is to classify the samples. That is, the fitness for individuals are evaluated by measure the classification accuracy for interaction of the features. To determine and update the fitness for each individual, it is possible to use any classification algorithm. But we introduce a gradient descendant rule for training data $\mathbf{D}$ as follows:

$$w_i = w_i + \eta(t_j - f(\mathbf{D}_j))m_{ji}, \tag{6.7}$$

where $w_i$ is a weight value for $i$-th feature and $t_j$ is a target class in the $j$-th training instance $\mathbf{D}_j$. $\eta$ is a learning rate and $m_{ji}$ is a value of the $i$-th attribute in the $j$-th instance. $f(\mathbf{D}_j)$ is a predicted output value of the $j$-th training instance by our model and determined as follows:

$$f(\mathbf{D}_j) = \begin{cases} 1, & \text{if } \sum_{i=0}^{n} w_i \cdot m_{ji} > 0 \\ -1, & \text{else} \end{cases} \tag{6.8}$$

The difference between the predictions and the target values specified in the training sequence is used to represent the error of the current weight vector. The target function is optimized to minimize the classification error. The weight values are evaluated against a sequence of training samples and are updated to improve the classification accuracy. The weight update processes are repeated until it would be converged after the number of epoch.

Using the learning scheme, we find most informative individuals for classification, that is, their absolute value of their weights is large. In addition, since our purpose is to identify a DNA methylation module, it might be necessary to find it if the number of the used feature is small. Finally, the fitness function for the $k$-th individual $X^k$, $Fitness(X^k)$ is defined as follows:

$$Fitness(X^k) = Acc(X^k) - Order(X^k), \tag{6.9}$$

where $Acc(X^k)$ is classification accuracy for training datasets and $Order(X^k)$ denotes the number of methylation sites which is selected in the individual $X^k$.

### 6.2.4 Dataset

The high-throughput DNA methylation profiling of large genomic regions could be produced by both microarray and NGS technologies. We applied our approach to these two types of datasets. The microarray data was generated by Illumina Infinium 27k Human DNAmethylation BeadChip in 1,475 samples, for surveying of genome-wide DNA methylation profiles in breast cancer and normal samples (Zhuang et al., 2012). Sequence-based datasets were produced by MethylCap-seq in matched normal and colon cancer samples and collected at GSE39068 (Simmer et al., 2012). The normalization and preprocessing was carried out using the same approaches to Simmer's works (Simmer et al., 2012).

## 6.3 Results

### 6.3.1 DNA methylation modules associated to breast cancer

This analysis has been carried out based on DNA methylation profiling datasets which experimentally measured the methylation status using DNAMethylation BeadChip (Zhuang et al., 2012). We extracted data for DNA methylation profiles on chromosome 17 from breast cancer and normal samples, and applied our method to the dataset. Figure 6.2 shows learning curves in the evolutionary process. The fitness value is improved when the number of generation is increased. Since we introduced a term for the number of the methylation sites to find a individual with the shorter length, the number of orders were decreased at the learning process (Figure 6.2). After convergence, 6 sites were selected for the discrimination, and these 6 sites are related to genes, KIAA1267, CD79B, ALOX12, TMEM98, KRT19 and FOXJ1 (Table 6.1).

ALOX12 have a role in growth of breast cancer and its inhibition may be a

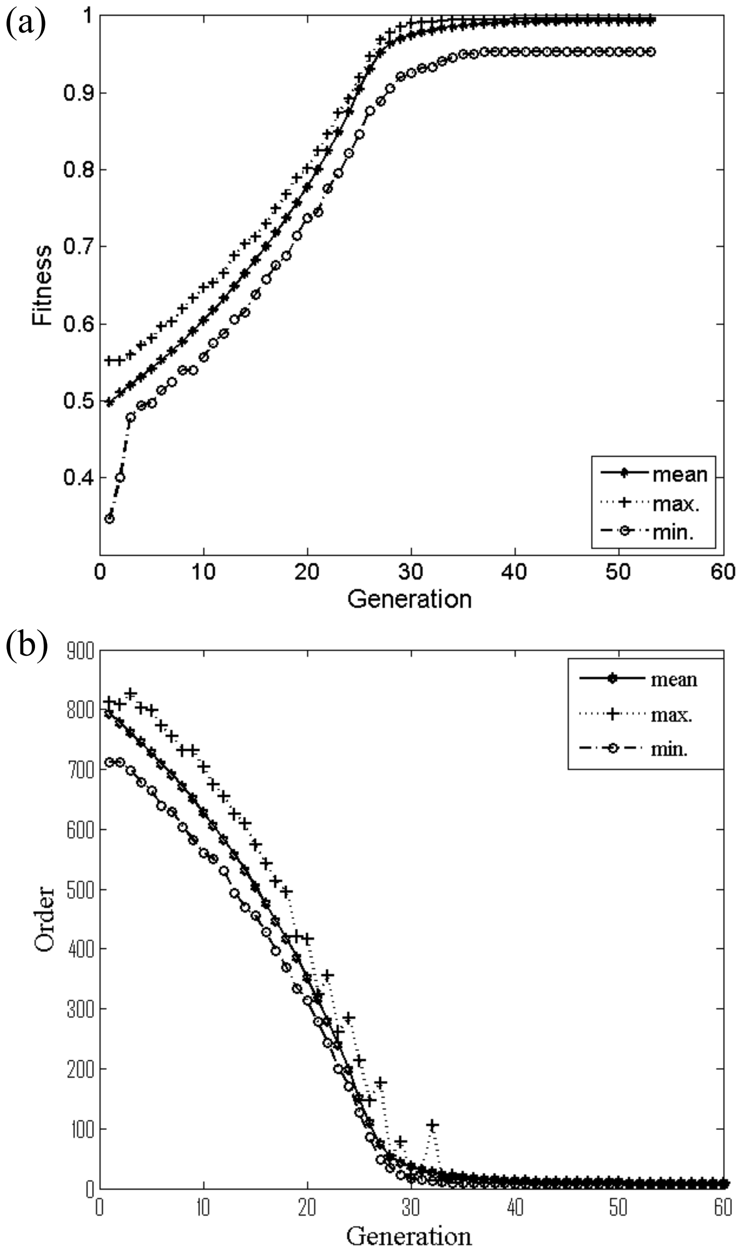


Figure 6.2: Learning curve using breast cancer datasets. x-axis is the number of generation and y-axis is (a) fitness values and (b) the number of orders.

Table 6.1: Finally selected methylation sites

| ID | Position | Gene | CGI location |
|---|---|---|---|
| cg02301815 | 41605268 | KIAA1267 | 41605074-41605445 |
| cg07973967 | 59363339 | CD79B | 25467633-25468370 |
| cg08946332 | 6840612 | ALOX12 | 6839463-6841283 |
| cg11833861 | 28279748 | TMEM98 | 28278827-28279833 |
| cg16585619 | 36938776 | KRT19 | NaN |
| cg24164563 | 71647990 | FOXJ1 | 71647419-71649480 |

strategy for inhibiting tumor growth (kumar Singh et al., 2012), the gene can be used as a serum marker for breast cancer (Singh et al., 2011). It is not clearly known how the ALOX12 methylation directly affects to breast cancer. However, it has been reported that hypermethylation of ALOX12 can be associated to cancer (Tan et al., 2009; Alvarez et al., 2010; Ammerpohl et al., 2012; Ohgami et al., 2012). Actually, the ALOX-12 gene is closely related to apoptosis, and the problem of the expression by the DNA methylation can cause a malfunction of the cell death (Ding et al., 1999; Pidgeon et al., 2002, 2003). Therefore, it might be reasonable that the change of methylation in the gene linked to most cancer, including breast tumor. KRT19 is a well-known marker for breast cancer patients (Ring et al., 2004; Lacroix, 2006), and KRT19 promoter is abberently methylated in cancer cell lines (Morris et al., 2008). Also, it has been reported that there exist the relationships between expression of CD79B and breast cancer (Ellsworth et al., 2008; Prat et al., 2010). FOXJ1, a member of the forkhead box (FOX) family, may function as a tumor suppressor gene in breast cancer (Jackson et al., 2010). FOXJ1 is hypermethylated and silenced in breast cancer cell lines (Demircan et al., 2009). TMEM98 is one of transmembrane

Table 6.2: Classification performance only using the 6 selected sites

| Algorithm | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Logistic Regression | 0.939 | 0.987 | 0.762 |
| SVM | 0.929 | 0.941 | 0.857 |
| Decision Tree | 0.939 | 0.952 | 0.867 |
| Naive Bayes | 0.919 | 0.951 | 0.765 |

proteins. Recently, Grimm et al. investigated the transmembrane proteins specific for cancer cells. The transmembrane protein can be a target for antibodies and be a biomarker for tumor diagnosis, prognosis, and treatment (Grimm et al., 2011). The function of KIAA1267 is not clearly known yet. But the gene encodes KAT8 regulatory NSL complex subunit 1, and the KAT8 regulates p53, a tumor suppressor gene (Li et al., 2009; Zhang et al., 2013). It imply the KIAA1267 can has a role in breast cancer.

Using the 6 sites, we tested classification performance using general machine learning algorithms (Table 6.2). To verify our method identified informative sites, we carried out classification only using the selected features. Table 6.2 shows the classification accuracy, sensitivity and specificity. Regardless the classification algorithms, it could be well-classified. For further verification, we randomly extracted the sites repeatedly (100 times), then measured the classification performance in each dataset. Figure 6.3 shows that the results of our method were higher than others, regardless of the number of the randomly selected sites.

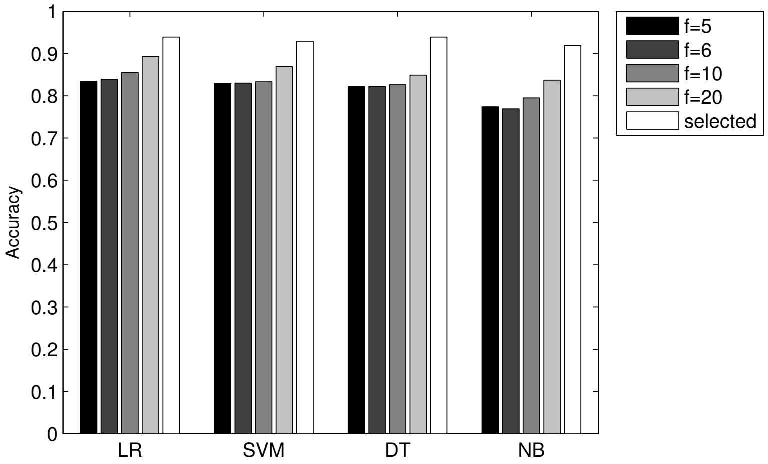


Figure 6.3: Classification accuracy using randomly selected sites. f is the number of the randomly selected sites, and white bar, marked as selected, is the results using only the 6 selected sites by our method. The results for the random datasets show averages of 100 times repeated experiments. LR: logistic regression, SVM: support vector machine, DT: decision tree, NB: naive Bayes.

### 6.3.2 Modules associated to colorectal cancer using high-throughput sequencing data

Recently, DNA methylation profiles could be measured by high-throughput sequencing technologies. We applied our method to the sequencing-based methylation profiling datasets produced by Simmer et al. (Simmer et al., 2012).

Figure 6.4 depicts improvement of the fitness in iterative learning procedures using these datasets. Among 10,393 genomic regions on chromosome 17 for the experiment, 348 regions were selected to discriminate the ovarian cancer and normal

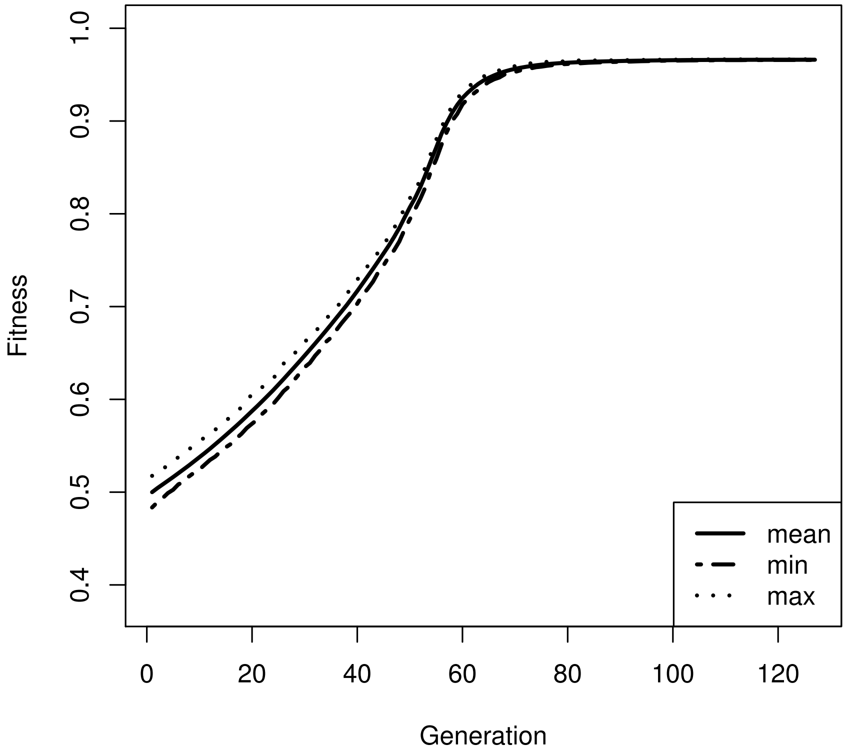


Figure 6.4: Learning curve using in colon cancer datasets. x-axis is the number of generation and y-axis is (a) fitness values.

samples after a convergence. Table 6.3 shows performance by classification algorithms using the 348 regions from the sequencing-based colorectal cancer datasets.

We annotated the selected regions using GPAT (Krebs et al., 2008) and investigated which genes were located closely on the selected regions. We accomplished gene set enrichment analysis (GSEA) with KEGG pathway using the genes whose

Table 6.3: Classification performance only using the 348 selected sites in colorectal cancer data

| Algorithm | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Logistic Regression | 0.900 | 0.920 | 0.880 |
| SVM | 0.940 | 0.960 | 0.920 |
| Decision Tree | 0.640 | 0.680 | 0.600 |
| Naive Bayes | 0.900 | 0.920 | 0.880 |

transcription start sites are located within 5000bp from the selected genomic regions. The GSEA was carried out using MSigDB (Subramanian et al., 2005; Liberzon et al., 2011). Table 6.4 summarizes the significantly enriched pathways with low p-values and shows that most of these are closely associated with cancer-related networks. Table 6.5 show the genes commonly enriched in the pathways. Note that the enriched signalling pathways were related to colorectal cancer. In colon cancer, the roles of wnt signalling pathway and MAPK signalling pathway have been very well-known (Jansson et al., 2005; Segditsas and Tomlinson, 2006; Fang and Richardson, 2005; Slattery et al., 2012). The genetic mutation affecting the pathway components and the alteration of their expression can enhance tumorigenicity in cancer cells. Also, neurotrophin signalling pathway could be related to growth of colorectal cancer cells (Akil et al., 2011) and chemokine signalling pathway suppresses colon cancer metastasis (Kitamura et al., 2010; Chen et al., 2012). Phosphatidylinositol signalling pathway plays an important role in the growth, survival and metabolism of cancer cells, and targeting this pathway has potential to lead to treatments for the colon cancer (Parsons et al., 2005; Yuan and Cantley, 2008). VEGF and ErbB can be valid therapeutic targets for patients with colon cancer (Ellis and Hicklin, 2008; Winder and Lenz, 2010; Roskoski Jr, 2004; Spano et al., 2005).

For further validation, we compared the results with ChIP-seq profiles of H3K4me3 and H3K27me3 at ENCODE project (Dunham et al., 2012). When we examined the selected sites on promoter regions, many of those were overlapped with the H3K4me3 and H3K27me3 binding sites with p-values of 1.86E-11 and 1.94E-05, respectively. The p-value for the regions overlapped with both of the two histone marks was 1.08E-05. The binding regions of the histone modification, called bivalent regions, were associated to cancer formation by abberant DNA methylation which leads to be silencing of regulators (Young et al., 2011; Chapman-Rothe et al., 2012). Since it is possible that DNA methylation are associated to bivalent regions in cancer, our studies would be help to understand the relationship between DNA methylation and chromatin signatures (McGarvey et al., 2008; Sharma et al., 2010; Balasubramanian et al., 2012; Reddington et al., 2013). Also it would help to investigate effects on cancer progression and possibilities for epigenomic treatments in cancer (Rodriguez et al., 2008; Mayor et al., 2011).

## 6.4 Discussion

DNA methylation can be also strongly associated with the complex diseases. It has been known that lots of genes are differentially methylated in various cancers or diseases. In this study, we presented a method to identify combinatorial effects of DNA methylation at multiple sites. From a systematic perspective, the relationship between DNA methylation regions and a specific disease is learned by the presented probabilistic evolutionary learning. The fitness value of a DNA methylation module measure the level of their responses to the disease. In computational view, our method can solve large scale problems by identifying modules with both compactness and high coverage of disease related genes. If the number of attributes is $n$, the number of possible cases is same to the number of elements in power set, $2^n$. Thus, the number of cases is exponentially increased according to the number of attributes. For example, if $n$=100, the number of cases is 1.27E30 and if $n$=1000, then the number is 1.07E301. However, Our method can find candidates in reasonable search in the problem space. In our every experiment, we found the candidate solution by searching less than 1.00E6 cases.

Applying our method to breast cancer and colorectal cancer data produced by high-throughput technologies, we detected the cancer-related modules confirmed by literatures and functional enrichment analysis. Interestingly we observed that the selected regions were located around genes which are enriched in cancer-related gene set categories significantly, and it provides evidence that the identified module in our study is biologically meaningful.

The studies on DNA methylation are likely to elucidate on the process of tumorigenesis as well as identify biomarkers. Our approaches which assist in the identification of multiple DNA methylation sites that have the potential to be epigenetically regulated might provide a useful strategy to detect epigenetic association related to a disease. The systematic identification of the disease-related genes and modules can provide insights into mechanisms underlying complex diseases and help efficient therapies or effective drugs.

By applying our method to microarray- and NGS-based data, we showed that it is applicable to a variety of data types and various disease contexts. Moreover, recent studies suggest that there exists a complex relationship between genetic variation, DNA methylation and so on. Systems genetic/epigenetics approaches are required for examining relationships among these. Although our framework is based on DNA methylation profiling datasets, it can be attempted to identify the combinatorial association for various factors including gene expression levels, microRNAs, copy number variation, genetic variations, and environmental factors. The integration of a variety of data would provide the basis for new hypothesis and experimental approaches in a model of complex disease.

In summary, we presented a method for searching the higher-order interaction of DNA methylation sites by a probabilistic evolutionary learning method. Using the approach, we also examined the potential for combined effects of various sites on genome. The results suggest that the alteration of DNA methylation at multiple sites affects on cancer. Similar to genome-wide association studies, our approach provides an opportunity to capture the complex and multifactorial relationship between the DNA methylation sites and to find new factors for future study. Therefore, our approach would be a way to facilitate a comprehensive analysis of genome-wide DNA methylation datasets and the interpretation for the effects of DNA methylation on multiple sites.

Table 6.4: Gene-set enrichment analysis annotated by promoter information using the 348 selected sites in colorectal cancer data

| Gene set | p-value | FDR q-value |
|---|---|---|
| Non-small cell lung cancer | 2.61E-05 | 4.25E-03 |
| Glioma | 4.56E-05 | 4.25E-03 |
| Neurotrophin signaling pathway | 3.25E-04 | 1.85E-02 |
| Pathways in cancer | 3.99E-04 | 1.85E-02 |
| Wnt signaling pathway | 5.52E-04 | 2.05E-02 |
| Aldosterone-regulated sodium reabsorption | 9.09E-04 | 2.22E-02 |
| Endocytosis | 9.62E-04 | 2.22E-02 |
| Vasopressin-regulated water reabsorption | 9.97E-04 | 2.22E-02 |
| Chemokine signaling pathway | 1.07E-03 | 2.22E-02 |
| Focal adhesion | 1.26E-03 | 2.34E-02 |
| Endometrial cancer | 1.39E-03 | 2.35E-02 |
| Basal cell carcinoma | 1.55E-03 | 2.41E-02 |
| Colorectal cancer | 1.97E-03 | 2.73E-02 |
| Pancreatic cancer | 2.50E-03 | 2.73E-02 |
| Melanoma | 2.57E-03 | 2.73E-02 |
| Chronic myeloid leukemia | 2.72E-03 | 2.73E-02 |
| Cytokine-cytokine receptor interaction | 2.82E-03 | 2.73E-02 |
| MAPK signaling pathway | 2.82E-03 | 2.73E-02 |
| Phosphatidylinositol signaling system | 2.94E-03 | 2.73E-02 |
| VEGF signaling pathway | 2.94E-03 | 2.73E-02 |
| Fc epsilon RI signaling pathway | 3.17E-03 | 2.81E-02 |
| Small cell lung cancer | 3.58E-03 | 2.98E-02 |
| ErbB signaling pathway | 3.83E-03 | 2.98E-02 |
| Apoptosis | 3.92E-03 | 2.98E-02 |
| Prostate cancer | 4.01E-03 | 2.98E-02 |

Table 6.5: Genes enriched in pathway analysis

| Gene Symbol | Description |
|---|---|
| TP53 | tumor protein p53 |
| PIK3R5 | phosphoinositide-3-kinase, regulatory subunit 5, p101 |
| PRKCA | protein kinase C, alpha |
| ARHGDIA | Rho GDP dissociation inhibitor (GDI) alpha |
| FZD2 | frizzled homolog 2 (Drosophila) |
| RABEP1 | rabaptin, RAB GTPase binding effector protein 1 |
| CCL16 | chemokine (C-C motif) ligand 16 |
| CXCL16 | chemokine (C-X-C motif) ligand 16 |
| CSF3 | colony stimulating factor 3 (granulocyte) |
| DUSP3 | dual specificity phosphatase 3 |
| ARSG | arylsulfatase G |

# Chapter 7

# Conclusion

Recently, explosive growth in data produced from various areas is continuously increasing. Intuitively the large amount of stored data contains valuable hidden knowledge, such that it could be used to improve the decision making process of an organization. There exists a clear need for the systematic methods for extracting the valuable knowledge from real-world datasets. This need has led to the emergence of a field called data mining and knowledge discovery. In order to extract or mine the knowledge or pattern of interest from data, intelligent mining tools are applied. The examples are association rule mining, clustering, classification, and so on.

Data collected from various biological domains is also becoming increasingly high in recent time. In particular, the large repositories of genome-wide measurement data pose the research question of how to retrieve valuable knowledge. In this dissertation, we proposed methods to identify higher-order interaction in genomic/epigenomic studies. We developed machine learning methods with evolutionary computation for extracting valuable information from large, high-dimensional data sets.

Statistical learning and evolutionary computation can be an way to mine the meaningful information from the biological big data. Especially, evolutionary computation has advantages to deal with a huge amount of the heterogeneous biological data. It appears to be more efficient in finding acceptable solutions than other random or semi-random search methods. Moreover, the approaches can be easily run in parallel, and allow groups of processor to be utilized for a search in the big data.

Furthermore, it might be helpful to exploit additional data sets even if they are only partially relevant for the data set of interest. For example, to further comprehensive understanding complex disease, it needs for integrative studies of various genomic and epigenomic datasets with environmental factors (Aschard et al., 2012). One advantage of our evolutionary machine learning approach is that it can easily extend and generalize the learning paradigms for multiple views of datasets. By systematically linking the various data sets, we would increases a chance to clarify biological knowledge and novel possibilities for biological results.

# Bibliography

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

Akil, H., Perraud, A., Mélin, C., Jauberteau, M.-O., and Mathonnet, M. (2011). Fine-tuning roles of endogenous brain-derived neurotrophic factor, trkb and sortilin in colorectal cancer cell survival. *PloS One*, 6(9):e25097.

Alvarez, S., Suela, J., Valencia, A., Fernández, A., Wunderlich, M., Agirre, X., Prósper, F., Martín-Subero, J. I., Maiques, A., Acquadro, F., et al. (2010). Dna methylation profiles and their relationship with cytogenetic status in adult acute myeloid leukemia. *PloS One*, 5(8):e12197.

Ammerpohl, O., Pratschke, J., Schafmayer, C., Haake, A., Faber, W., von Kampen, O., Brosch, M., Sipos, B., von Schönfels, W., Balschun, K., et al. (2012). Distinct dna methylation patterns in cirrhotic liver and hepatocellular carcinoma. *Int. J. Cancer*, 130(6):1319–1328.

Aschard, H., Lutz, S., Maus, B., Duell, E. J., Fingerlin, T. E., Chatterjee, N., Kraft, P., and Van Steen, K. (2012). Challenges and opportunities in genome-wide environmental interaction (gwei) studies. *Hum. Genet.*, 131(10):1591–1613.

Bach, F. R. and Jordan, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48.

Bailey, T. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings International Conference on Intelligent Systems for Molecular Biology; ISMB.*, volume 2, page 28.

Balasubramanian, D., Akhtar-Zaidi, B., Song, L., Bartels, C. F., Veigl, M., Beard, L., Myeroff, L., Guda, K., Lutterbaugh, J., Willis, J., et al. (2012). H3k4me3 inversely correlates with dna methylation at a large class of non-cpg-island-containing start sites. *Genome Med.*, 4(5):47.

Banerjee, N. and Zhang, M. Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, 31(23):7024–7031.

Baylin, S. B. and Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer*, 11(10):726–734.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., Pritchard, J. K., et al. (2011). Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol.*, 12(1):R10.

Bernado-Mansilla, E. and Garrell, J. (2003). Accuracy-based learning classifier systems: models, analysis and applications to classification tasks. *Evol. Comput.*, 11(3):209–238.

Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128(4):669–681.

Bloushtain-Qimron, N., Yao, J., Snyder, E. L., Shipitsin, M., Campbell, L. L., Mani, S. A., Hu, M., Chen, H., Ustyansky, V., Antosiewicz, J. E., et al. (2008). Cell type-specific dna methylation patterns in the human breast. *Proc. Natl. Acad. Sci. U. S. A.*, 105(37):14076–14081.

Bonetta, L. (2008). Epigenomics: Detailed analysis. *Nature*, 454:795–98.

Brāzma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8(11):1202–1215.

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D., and Scandura, J. M. (2011). Dna methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One*, 6(1):e14524.

Brinkman, A. B., Simmer, F., Ma, K., Kaan, A., Zhu, J., and Stunnenberg, H. G. (2010). Whole-genome dna methylation profiling using methylcap-seq. *Methods*, 52(3):232–236.

Brown, R. L., Reinke, L. M., Damerow, M. S., Perez, D., Chodosh, L. A., Yang, J., and Cheng, C. (2011). Cd44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J. Clin. Invest.*, 121(3):1064.

Bush, W. and Moore, J. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, 8:12.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.*, 27(2):167–174.

Ceberio, J., Irurozqui, E., Mendiburu, A., and Lozano, J. (2013). A distance-based ranking model estimation of distribution algorithm for the flowshop scheduling problem. *IEEE Trans. Evol. Comput.*, PP(99):1–1.

Chapman-Rothe, N., Curry, E., Zeller, C., Liber, D., Stronach, E., Gabra, H., Ghaem-Maghami, S., and Brown, R. (2012). Chromatin h3k27me3/h3k4me3 histone marks define gene sets in high-grade serous ovarian cancer that distinguish malignant, tumour-sustaining and chemo-resistant ovarian tumour cells. *Oncogene*.

Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R., and Adjaye, J. (2010). Computational analysis of genome-wide dna methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res.*, 20(10):1441–1450.

Chen, H. J., Edwards, R., Tucci, S., Bu, P., Milsom, J., Lee, S., Edelmann, W., Gümüs, Z. H., Shen, X., and Lipkin, S. (2012). Chemokine 25–induced signaling suppresses colon cancer invasion and metastasis. *J. Clin. Invest.*, 122(9):3184.

Chen, T., Lehre, P., Tang, K., and Yao, X. (2009). When is an estimation of distribution algorithm better than an evolutionary algorithm? In *Evolutionary Computation, 2009. CEC '09. IEEE Congress on*, pages 1470–1477.

Cheung, H., Lee, T., Davis, A., Taft, D., Rennert, O., and Chan, W. (2010). Genome-wide dna methylation profiling reveals novel epigenetically regulated genes and non-coding rnas in human testicular cancer. *Brit. J. Cancer*, 102(2):419–427.

Chin, L., Hahn, W. C., Getz, G., and Meyerson, M. (2011). Making sense of cancer genomic data. *Genes Dev.*, 25(6):534–555.

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H.-J., Yoon, D., Lee, M. H., Kim, D.-J., Park, M., et al. (2009). A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.*, 41:527–534.

Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. *Annu. Rev. Cell Dev.Biol.*, 26:721.

Cieply, B., Riley, P., Pifer, P. M., Widmeyer, J., Addison, J. B., Ivanov, A. V., Denvir, J., and Frisch, S. M. (2012). Suppression of the epithelial–mesenchymal transition by grainyhead-like-2. *Cancer Res.*, 72(9):2440–2453.

Clark, T., De Iorio, M., and Griffths, R. (2008). An evolutionary algorithm to find associations in dense genetic maps. *IEEE Trans. Evol. Comput.*, 12(3):297–306.

Cooper, L. R., Corne, D. W., and Crabbe, M. J. C. (2003). Use of a novel hill-climbing genetic algorithm in protein folding simulations. *Comput. Biol. Chem.*, 27(6):575–580.

Cordell, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, 10(6):392–404.

Das, D., Banerjee, N., and Zhang, M. Q. (2004). Interacting models of cooperative gene regulation. *Proc. Natl. Acad. Sci. U. S. A.*, 101(46):16234–16239.

Das, M. and Dai, H.-K. (2007). A survey of dna motif finding algorithms. *BMC Bioinformatics*, 8(Suppl 7):S21.

Deb, K. and Datta, R. (2010). A fast and accurate solution of constrained optimization problems using a hybrid bi-objective and penalty function approach. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–8.

Demircan, B., Dyer, L. M., Gerace, M., Lobenhofer, E. K., Robertson, K. D., and Brown, K. D. (2009). Comparative epigenomics of human and mouse mammary tumors. *Genes, Chromosomes and Cancer*, 48(1):83–97.

Ding, X.-Z., Kuszynski, C. A., El-Metwally, T. H., and Adrian, T. E. (1999). Lipoxygenase inhibition induced apoptosis, morphological changes, and carbonic anhydrase expression in human pancreatic cancer cells. *Biochem. Biophys. Res. Commun.*, 266(2):392–399.

Dohrmann, P. R., Butler, G., Tamai, K., Dorland, S., Greene, J. R., Thiele, D. J., and Stillman, D. J. (1992). Parallel pathways of gene regulation: homologous regulators swi5 and ace2 differentially control transcription of ho and chitinase. *Genes Dev.*, 6(1):93–104.

Dohrmann, P. R., Voth, W. P., and Stillman, D. J. (1996). Role of negative regulation in promoter specificity of the homologous transcriptional activators ace2p and swi5p. *Mol. Cell. Biol.*, 16(4):1746–1758.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–731.

Dumont, N., Wilson, M. B., Crawford, Y. G., Reynolds, P. A., Sigaroudinia, M., and Tlsty, T. D. (2008). Sustained induction of epithelial to mesenchymal transition activates dna methylation of genes silenced in basal-like breast cancers. *Proc. Natl. Acad. Sci. U. S. A.*, 105(39):14867–14872.

Dunham, I., Kundaje, A., Aldred, S., Collins, P., Davis, C., Doyle, F., Epstein, C., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Easwaran, H., Johnstone, S., Van Neste, L., Ohm, J., Mosbruger, T., Wang, Q., Aryee, M., Joyce, P., Ahuja, N., Weisenberger, D., Collisson, E., Zhu, J., Yegnasubramanian, S., Matsui, W., and Baylin, S. (2012). A dna hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Res.*, 22:837–849.

Ellis, L. M. and Hicklin, D. J. (2008). Vegf-targeted therapy: mechanisms of anti-tumour activity. *Nat. Rev. Cancer*, 8(8):579–591.

Ellsworth, R., Heckman, C., Seebach, J., Field, L., Love, B., Hooke, J., and Shriver, C. (2008). Identification of a gene expression breast cancer nodal metastasis profile. In *J. Clin. Oncol. (Meeting Abstracts)*, volume 26, page 1022.

Esteller, M. (2007). Epigenetic gene silencing in cancer: the dna hypermethylome. *Hum. Mol. Genet.*, 16(R1):R50–R59.

Esteller, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.*, 358(11):1148–1159.

Fadare, O. and Tavassoli, F. A. (2008). Clinical and pathologic aspects of basal-like breast cancers. *Nat. Clin. Pract. Oncol.*, 5(3):149–159.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G., Shen, R., Seshan, V., Mo, Q., Heguy, A., et al. (2011). Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci. Transl. Med.*, 3(75):75ra25.

Fang, J. Y. and Richardson, B. C. (2005). The mapk signalling pathways and colorectal cancer. *Lancet Oncol.*, 6(5):322–327.

Feero, W. G., Guttmacher, A. E., and Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, 363(2):166–176.

Fernandez, A., Garcia, S., Luengo, J., Bernado-Mansilla, E., and Herrera, F. (2010). Genetics-based machine learning for rule induction: State of the art, taxonomy, and comparative study. *IEEE Trans. Evol. Comput.*, 14(6):913–941.

Fiaschi, L., Garibaldi, J., and Krasnogora, N. (2010). A framework for the application of decision trees to the analysis of snps data. In *IEEE Symposium on*

*Computational Intelligence in Bioinformatics and Computational Biology 2009 (CIBCB 09)*, pages 106–113.

Fogel, G. B. and Corne, D. W. (2002). *Evolutionary computation in bioinformatics*. Morgan Kaufmann.

Fratkin, E., Naughton, B. T., Brutlag, D. L., and Batzoglou, S. (2006). Motifcut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics*, 22(14):e150–e157.

Grimm, D., Bauer, J., Pietsch, J., Infanger, M., Eucker, J., Eilles, C., and Schoenberger, J. (2011). Diagnostic and therapeutic use of membrane proteins in cancer cells. *Curr. Med. Chem.*, 18(2):176–190.

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

Handel, A. E., Ebers, G. C., and Ramagopalan, S. V. (2010). Epigenetics: molecular mechanisms and implications for disease. *Trends Mol. Med.*, 16(1):7–16.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664.

Heap, G. A., Trynka, G., Jansen, R. C., Bruinenberg, M., Swertz, M. A., Dinesen, L. C., Hunt, K. A., Wijmenga, C., Franke, L., et al. (2009). Complex nature of snp genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics*, 2(1):1.

Heidema, A., Feskens, E., Doevendans, P., Ruven, H., van Houwelingen, H., Mariman, E., and Boer, J. (2007). Analysis of multiple snps in genetic association studies: comparison of three multi-locus methods to prioritize and select snps. *Genet Epidemiol.*, 31(8):910–921.

Hill, V. K., Ricketts, C., Bieche, I., Vacher, S., Gentle, D., Lewis, C., Maher, E. R., and Latif, F. (2011). Genome-wide dna methylation profiling of cpg islands in breast cancer identifies novel genes associated with tumorigenicity. *Cancer Res.*, 71(8):2988–2999.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6(2):95–108.

Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80.

Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jönsson, G., Olsson, H., Borg, Å., and Ringnér, M. (2010). Molecular subtypes of breast cancer are associated with characteristic dna methylation patterns. *Breast Cancer Res.*, 12(3):R36.

Hong, C., Kim, Y. J., Moon, S., Shin, Y.-A., Cho, Y. S., and Lee, J.-Y. (2012). Karebrowser: Snp database of korea association resource project. *BMB Rep.*, 45(1):47–50.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R., Boks, M., van Eijk, K., van den Berg, L., and Ophoff, R. (2012). Aging effects on dna methylation modules in human brain and blood tissue. *Genome Biol.*, 13(10):R97.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of¡ i¿ cis¡/i¿-regulatory elements associated with groups of functionally related genes in¡ i¿ saccharomyces cerevisiae¡/i¿. *J. Mol. Biol.*, 296(5):1205–1214.

Huttenhower, C. and Hofmann, O. (2010). A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779.

Hvidsten, T. R., Wilczyński, B., Kryshtafovych, A., Tiuryn, J., Komorowski, J., and Fidelis, K. (2005). Discovering regulatory binding-site modules using rule-based learning. *Genome Res.*, 15(6):856–866.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nature Genet.*, 41(2):178–186.

Jackson, B. C., Carpenter, C., Nebert, D. W., Vasiliou, V., et al. (2010). Update of human and mouse forkhead box (fox) gene families. *Hum. Genomics*, 4:345–352.

Janssens, A. and van Duijn, C. (2008). Genome-based prediction of common diseases: advances and prospects. *Hum. Mol. Genet.*, 17(R2):R166–R173.

Jansson, E. Å., Are, A., Greicius, G., Kuo, I.-C., Kelly, D., Arulampalam, V., and Pettersson, S. (2005). The wnt/$\beta$-catenin signaling pathway targets ppar$\gamma$ activity in colon cancer cells. *Proc. Natl. Acad. Sci. U. S. A.*, 102(5):1460–1465.

Jeffery, I. B., Madden, S. F., McGettigan, P. A., Perriere, G., Culhane, A. C., and Higgins, D. G. (2007). Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, 23(3):298–305.

Jones, P. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492.

Jones, P. A. and Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4):683–962.

Jones, P. A. and Takai, D. (2001). The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070.

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.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489.

Kasturi, J. and Acharya, R. (2005). Clustering of diverse genomic data using information fusion. *Bioinformatics*, 21(4):423–429.

Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). Match: a tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res.*, 31(13):3576–3579.

Keleher, C. A., Passmore, S., and Johnson, A. (1989). Yeast repressor alpha 2 binds to its operator cooperatively with yeast protein mcm1. *Mol. Cell. Biol.*, 9(11):5228–5230.

Keles, S., van der Laan, M. J., and Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–1175.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. A., Niveleau, A., Cedar, H., et al. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.*, 38(2):149–153.

Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. F., Moran, J. L., et al. (2012). Exome sequencing and the genetic basis of complex traits. *Nat. Genet.*, 44(6):623–630.

Kim, J. H., Dhanasekaran, S. M., Prensner, J. R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M., et al. (2011). Deep sequencing reveals distinct patterns of dna methylation in prostate cancer. *Genome Res.*, 21(7):1028–1041.

Kim, S., Cho, H., Lee, D., and Webster, M. (2012). Association between snps and gene expression in multiple regions of the human brain. *Translational Psychiatry*, 2(5):e113.

Kitamura, T., Fujishita, T., Loetscher, P., Revesz, L., Hashida, H., Kizaka-Kondoh, S., Aoki, M., and Taketo, M. M. (2010). Inactivation of chemokine (cc motif) receptor 1 (ccr1) suppresses colon cancer liver metastasis by blocking accumulation of immature myeloid cells in a mouse model. *Proc. Natl. Acad. Sci. U. S. A.*, 107(29):13063–13068.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70.

Kouskoumvekaki, I., Shublaq, N., and Brunak, S. (2013). Facilitating the use of large-scale biological data and tools in the era of translational bioinformatics. *Brief. Bioinform.*, page bbt055.

Krebs, A., Frontini, M., and Tora, L. (2008). Gpat: retrieval of genomic annotation from large genomic position datasets. *BMC Bioinformatics*, 9(1):533.

Kumar, M., Husian, M., Upreti, N., and Gupta, D. (2010). Genetic algorithm: Review and application. *Int. J. Inform. Tech. Knowl. Manag.*, 2(2):451–454.

kumar Singh, A., Singh, R., Naz, F., Chauhan, S. S., Dinda, A., Shukla, A. A., Gill, K., Kapoor, V., and Dey, S. (2012). Structure based design and synthesis of peptide inhibitor of human lox-12: in vitro and in vivo analysis of a novel therapeutic agent for breast cancer. *PloS One*, 7(2):e32521.

Lacroix, M. (2006). Significance, detection and markers of disseminated breast cancer cells. *Endocr.-Relat. Cancer*, 13(4):1033–1067.

Laird, P. (2010). Principles and challenges of genomewide dna methylation analysis. *Nat. Rev. Genet.*, 11(3):191–203.

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, 274(5287):536–539.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Li, X., Wu, L., Corsa, C. A. S., Kunkel, S., and Dou, Y. (2009). Two mammalian mof complexes regulate transcription activation by distinct mechanisms. *Mol. Cell*, 36(2):290–301.

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

Liu, F., Tsai, J.-P., Chen, R., Chen, S., and Shih, S. H. (2004). Fmga: finding motifs by genetic algorithm. In *Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on*, pages 459–466.

Lombaerts, M., Van Wezel, T., Philippo, K., Dierssen, J., Zimmerman, R., Oosting, J., van Eijk, R., Eilers, P., van De Water, B., Cornelisse, C., et al. (2006). E-cadherin transcriptional downregulation by promoter methylation but not mutation is related to epithelial-to-mesenchymal transition in breast cancer cell lines. *Brit. J. Cancer*, 94(5):661–671.

Lydall, D., Ammerer, G., and Nasmyth, K. (1991). A new role for mcm1 in yeast: cell cycle regulation of sw15 transcription. *Genes Dev.*, 5(12b):2405–2419.

MacKay, V. L., Mai, B., Waters, L., and Breeden, L. L. (2001). Early cell cycle box-mediated transcription ofcln3 and swi4 contributes to the proper timing of the g1-to-s transition in budding yeast. *Mol. Cell. Biol.*, 21(13):4140–4148.

Mahony, S., Hendrix, D., Golden, A., Smith, T. J., and Rokhsar, D. S. (2005). Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–1814.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., et al. (2003). Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34(suppl 1):D108–D110.

Mayor, R., Muñoz, M., Coolen, M. W., Custodio, J., Esteller, M., Clark, S. J., and Peinado, M. A. (2011). Dynamics of bivalent chromatin domains upon drug induced reactivation and resilencing in cancer cells. *Epigenetics*, 6(9):1138–1148.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, 9(5):356–369.

McGarvey, K. M., Van Neste, L., Cope, L., Ohm, J. E., Herman, J. G., Van Criekinge, W., Schuebel, K. E., and Baylin, S. B. (2008). Defining a chromatin pattern that characterizes dna-hypermethylated genes in colon cancer cells. *Cancer Res.*, 68(14):5753–5759.

Mead, J., Zhong, H., Acton, T. B., and Vershon, A. K. (1996). The yeast alpha2 and mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. *Mol. Cell. Biol.*, 16(5):2135–2143.

Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., Baylin, S. B., and Sidransky, D. (1995). 52 cpg island methylation is associated with transcriptional silencing of the tumour suppressor p16/cdkn2/mts1 in human cancers. *Nat. Med.*, 1(7):686–692.

Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nat. Rev. Genet.*, 11(1):31–46.

Moore, J., Asselbergs, F., and Williams, S. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455.

Moore, J. and White, B. (2007). *Genome-wide association studies for common dis-*

*eases and complex traits.* Genetic Programming Theory and Practice IV. Springer, New York.

Morillon, A., O'Sullivan, J., Azad, A., Proudfoot, N., and Mellor, J. (2003). Regulation of elongating rna polymerase ii by forkhead transcription factors in yeast. *Science*, 300(5618):492–495.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662.

Morris, M., Gentle, D., Abdulrahman, M., Clarke, N., Brown, M., Kishida, T., Yao, M., Teh, B., Latif, F., and Maher, E. R. (2008). Functional epigenomics approach to identify methylated candidate tumour suppressor genes in renal cell carcinoma. *Br. J. Cancer*, 98(2):496–501.

Morrow, B. E., Johnson, S. P., and Warner, J. R. (1989). Proteins that bind to the yeast rdna enhancer. *J. Biol. Chem.*, 264(15):9061–9068.

Murrell, A., Rakyan, V. K., and Beck, S. (2005). From genome to epigenome. *Hum. Mol. Genet.*, 14(suppl 1):R3–R10.

Namkung, J., Nam, J., and Park, T. (2007). Identification of expression quantitative trait loci by the interaction analysis using genetic algorithm. *BMC Proc.*, 1(1):S69.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527.

Nguyen, H. D., Yoshihara, I., Yamamori, K., and Yasunaga, M. (2002). A parallel hybrid genetic algorithm for multiple protein sequence alignment. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, pages 309–314. IEEE.

Notredame, C. and Higgins, D. G. (1996). Saga: sequence alignment by genetic algorithm. *Epigenetics*, 24(8):1515–1524.

Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K., and Wegener, I. (2007). Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 23(24):3280–3288.

Ohgami, R. S., Ma, L., Ren, L., Weinberg, O. K., Seetharam, M., Gotlib, J. R., and Arber, D. A. (2012). Dna methylation analysis of alox12 and gstm1 in acute myeloid leukaemia identifies prognostically significant groups. *Brit. J. Haematol.*, 159(2):182–190.

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., and Gilad, Y. (2011). A genome-wide study of dna methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.*, 7(2):e1001316.

Pal, S., Bandyopadhyay, S., and Ray, S. (2006). Evolutionary computation in bioinformatics: a review. *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, 36(5):601–615.

Palsson, B. and Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.*, 6(11):787–789.

Park, P. J., Butte, A. J., and Kohane, I. S. (2002). Comparing expression profiles of genes with similar promoter regions. *Bioinformatics*, 18(12):1576–1584.

Parsons, D. W., Wang, T.-L., Samuels, Y., Bardelli, A., Cummins, J. M., DeLong, L., Silliman, N., Ptak, J., Szabo, S., Willson, J. K., et al. (2005). Colorectal cancer: mutations in a signalling pathway. *Nature*, 436(7052):792–792.

Pelikan, M. (2006). Implementation of the dependency-tree estimation of distribution algorithm in c++.

Pelikan, M., Tsutsui, S., and Kalapala, R. (2007). Dependency trees, permutations, and quadratic assignment problem. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, GECCO '07, pages 629–629, New York, NY, USA. ACM.

Pidgeon, G. P., Kandouz, M., Meram, A., and Honn, K. V. (2002). Mechanisms controlling cell cycle arrest and induction of apoptosis after 12-lipoxygenase inhibition in prostate cancer cells. *Cancer Res.*, 62(9):2721–2727.

Pidgeon, G. P., Tang, K., Cai, Y. L., Piasentin, E., and Honn, K. V. (2003). Overexpression of platelet-type 12-lipoxygenase promotes tumor cell survival by enhancing $\alpha$v$\beta$3 and $\alpha$v$\beta$5 integrin expression. *Cancer Res.*, 63(14):4258–4267.

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29(2):153–159.

Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. *Nat. Biotech.*, 28(10):1057–1068.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.*, 12(5):R68.

Rao, X., Evans, J., Chae, H., Pilrose, J., Kim, S., Yan, P., Huang, R., Lai, H., Lin, H., Liu, Y., et al. (2013). Cpg island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, 32:4519—-4528.

Rauch, T. A. and Pfeifer, G. P. (2010). Dna methylation profiling using the methylated-cpg island recovery assay (mira). *Methods*, 52(3):213–217.

Reddington, J. P., Perricone, S. M., Nestor, C. E., Reichmann, J., Youngson, N. A., Suzuki, M., Reinhardt, D., Dunican, D. S., Prendegast, J. G., Mjoseng, H., et al. (2013). Redistribution of h3k27me3 upon dna hypomethylation results in derepression of polycomb-target genes. *Genome Biol.*, 14(3):R25.

Rhee, J.-K., Joung, J.-G., Chang, J.-H., Fei, Z., and Zhang, B.-T. (2009). Identification of cell cycle-related regulatory motifs using a kernel canonical correlation analysis. *BMC Genomics*, 10(Suppl 3):S29.

Rhee, J.-K., Kim, K., Chae, H., Evans, J., Yan, P., Zhang, B.-T., Gray, J., Spellman, P., Huang, T. H.-M., Nephew, K. P., et al. (2013). Integrated analysis of genome-wide dna methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Res.*, 41(18):8464–8474.

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T. R., Ghosh, D., and Chinnaiyan, A. M. (2005). Mining for regulatory programs in the cancer transcriptome. *Nat. Genet.*, 37(6):579–583.

Ring, A., Smith, I. E., and Dowsett, M. (2004). Circulating tumour cells in breast cancer. *Lancet Oncol.*, 5(2):79–88.

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F., and Moore, J. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, 69(1):138–147.

Robertson, K. D. (2005). Dna methylation and human disease. *Nat. Rev. Genet.*, 6(8):597–610.

Robinson, M. D., Stirzaker, C., Statham, A. L., Coolen, M. W., Song, J. Z., Nair, S. S., Strbenac, D., Speed, T. P., and Clark, S. J. (2010). Evaluation of affinity-based genome-wide dna methylation data: effects of cpg density, amplification bias, and copy number variation. *Genome Res.*, 20(12):1719–1729.

Rodriguez, J., Muñoz, M., Vives, L., Frangou, C. G., Groudine, M., and Peinado, M. A. (2008). Bivalent domains enforce transcriptional memory of dna methylated genes in cancer cells. *Proc. Natl. Acad. Sci. U. S. A.*, 105(50):19809–19814.

Roskoski Jr, R. (2004). The erbb/her receptor protein-tyrosine kinases and cancer. *Biochem. Biophys. Res. Commun.*, 319(1):1–11.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2004). Exploring interactions in high dimensional genomic data: An overview of logic regression, with applications. *J. Multivar. Anal.*, 90:178–195.

Ruike, Y., Imanaka, Y., Sato, F., Shimizu, K., and Tsujimoto, G. (2010). Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-dna immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, 11(1):137.

Sadikovic, B., Al-Romaih, K., Squire, J., and Zielenska, M. (2008). Cause and consequences of genetic and epigenetic alterations in human cancer. *Curr. Genomics*, 9(6):394–408.

Sakai, T., Toguchida, J., Ohtani, N., Yandell, D. W., Rapaport, J. M., and Dryja, T. P. (1991). Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am. J. Hum. Genet.*, 48(5):880.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32(suppl 1):D91–D94.

Sandoval, J. and Esteller, M. (2012). Cancer epigenomics: beyond genomics. *Curr. Opin. Genet. Dev.*, 22(1):50–55.

Santana, R., Mendiburu, A., Zaitlen, N., Eskin, E., and Lozano, J. (2010). Multimarker tagging single nucleotide polymorphism selection using estimation of distribution algorithms. *Artif. Intell. Med.*, 50(3):193–201.

Savage, K., Lambros, M. B., Robertson, D., Jones, R. L., Jones, C., Mackay, A., James, M., Hornick, J. L., Pereira, E. M., Milanezi, F., et al. (2007). Caveolin 1 is overexpressed and amplified in a subset of basal-like and metaplastic breast carcinomas: a morphologic, ultrastructural, immunohistochemical, and in situ hybridization analysis. *Clin. Cancer Res.*, 13(1):90–101.

Segal, E., Yelensky, R., and Koller, D. (2003). Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19(suppl 1):i273–i282.

Segditsas, S. and Tomlinson, I. (2006). Colorectal cancer and genetic alterations in the wnt pathway. *Oncogene*, 25(57):7531–7537.

Sharma, S., Kelly, T. K., and Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36.

Shelke, K., Jayaraman, S., Ghosh, S., and Valadi, J. (2013). Hybrid feature selection and peptide binding affinity prediction using an eda based algorithm. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 2384–2389.

Shim, V., Tan, K., Chia, J., and Al Mamun, A. (2013). Multi-objective optimization with estimation of distribution algorithm in a noisy environment. *Evol. Comput.*, 21(1):149–177.

Sigaud, O. and Wilson, S. (2007). Learning classifier systems: a survey. *Soft Comput.*, 11(11):1065–1078.

Simmer, F., Brinkman, A., Assenov, Y., Matarese, F., Kaan, A., Sabatino, L., Villanueva, A., Huertas, D., Esteller, M., Lengauer, T., Bock, C., Colantuoni, V., Altucci, L., and Stunnenberg, H. (2012). Comparative genome-wide dna methylation analysis of colorectal tumor and matched normal tissues. *Epigenetics*, 7(12):1355–1367.

Simon, D. (2013). *Evolutionary Optimization Algorithms: Biologically-Inspired and Population-Based Approaches to Computer Intelligence.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., et al. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708.

Singh, A., Kant, S., Parshad, R., Banerjee, N., and Dey, S. (2011). Evaluation of human lox-12 as a serum marker for breast cancer. *Biochem. Biophys. Res. Commun.*, 414(2):304–308.

Slattery, M. L., Lundgreen, A., and Wolff, R. K. (2012). Map kinase genes and colon and rectal cancer. *Carcinogenesis*, 33(12):2398–2408.

Sloan, E. K., Stanley, K. L., and Anderson, R. L. (2004). Caveolin-1 inhibits breast cancer growth and metastasis. *Oncogene*, 23(47):7893–7897.

Spano, J., Fagard, R., Soria, J.-C., Rixe, O., Khayat, D., and Milano, G. (2005). Epidermal growth factor receptor signaling in colorectal cancer: preclinical data and therapeutic perspectives. *Ann. Oncol.*, 16(2):189–194.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297.

Sproul, D., Nestor, C., Culley, J., Dickson, J. H., Dixon, J. M., Harrison, D. J., Meehan, R. R., Sims, A. H., and Ramsahoye, B. H. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 108(11):4364–4369.

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383.

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

Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R., Carr, J. M., Khrebtukova, I., Luo, S., Zhang, L., et al. (2011). Integrated analysis of gene expression, cpg island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, 6(2):e17490.

Suzuki, M. M. and Bird, A. (2008). Dna methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, 9(6):465–476.

Szymczak, S., Biernacka, J., Cordell, H., Gonzalez-Recio, O., Konig, I., Zhang, H., and Sun, Y. (2009). Machine learning in genome-wide association studies. *Genet. Epidemiol.*, 33(Suppl 1):S51–57.

Takai, D. and Jones, P. A. (2002). Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.*, 99(6):3740–3745.

Tan, A. C., Jimeno, A., Lin, S. H., Wheelhouse, J., Chan, F., Solomon, A., Rajeshkumar, N., Rubio-Viqueira, B., and Hidalgo, M. (2009). Characterizing dna methylation patterns in pancreatic cancer genome. *Mol. Oncol.*, 3(5):425–438.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.*, 22(3):281–285.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E., and Schork, N. (2011). The importance of phase information for human genomics. *Nat. Rev. Genet.*, 12(3):213–223.

Thiery, J. P. (2002). Epithelial–mesenchymal transitions in tumour progression. *Nat. Rev. Cancer*, 2(6):442–454.

Thiery, J. P., Acloque, H., Huang, R. Y., and Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *Cell*, 139(5):871–890.

Toft, D. J. and Cryns, V. L. (2011). Minireview: basal-like breast cancer: from molecular profiles to targeted therapies. *Mol. Endocrinol.*, 25(2):199–211.

Toperoff, G., Aran, D., Kark, J. D., Rosenberg, M., Dubnikov, T., Nissan, B., Wainstein, J., Friedlander, Y., Levy-Lahad, E., Glaser, B., et al. (2012). Genome-wide survey reveals predisposing diabetes type 2-related dna methylation variations in human peripheral blood. *Hum. Mol. Genet.*, 21(2):371–383.

Tsai, H.-K., Lu, H. H.-S., and Li, W.-H. (2005). Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, 102(38):13532–13537.

Ueki, T., Walter, K. M., Skinner, H., Jaffee, E., Hruban, R. H., and Goggins, M. (2002). Aberrant cpg island methylation in cancer cell lines arises in the primary cancers from which they were derived. *Oncogene*, 21(13):2114–2117.

Unger, R. and Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, 231(1):75–81.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Vershon, A. K. and Johnson, A. D. (1993). A short, disordered protein region mediates interactions between the homeodomain of the yeast $\alpha 2$ protein and the mcm1 protein. *Cell*, 72(1):105–112.

Walker, B. A., Wardell, C. P., Chiecchio, L., Smith, E. M., Boyd, K. D., Neri, A., Davies, F. E., Ross, F. M., and Morgan, G. J. (2011). Aberrant global methylation patterns affect the molecular pathogenesis and prognosis of multiple myeloma. *Blood*, 117(2):553–562.

Wang, B., Chen, P., Zhang, J., Zhao, G., and Zhang, X. (2010). Inferring protein-protein interactions using a hybrid genetic algorithm/support vector machine method. *Protein Pept. Lett.*, 17(9):1079–1084.

Wang, R., Purshouse, R. C., and Fleming, P. J. (2013). On finding well-spread pareto optimal solutions by preference-inspired co-evolutionary algorithm. In *Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference*, GECCO '13, pages 695–702, New York, NY, USA. ACM.

Werth, M., Walentin, K., Aue, A., Schönheit, J., Wuebken, A., Pode-Shakked, N., Vilianovitch, L., Erdmann, B., Dekel, B., Bader, M., et al. (2010). The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex. *Development*, 137(22):3835–3845.

Winder, T. and Lenz, H.-J. (2010). Vascular endothelial growth factor and epidermal growth factor signaling pathways as therapeutic targets for colorectal cancer. *Gastroenterology*, 138(6):2163–2176.

Xiang, X., Deng, Z., Zhuang, X., Ju, S., Mu, J., Jiang, H., Zhang, L., Yan, J., Miller, D., and Zhang, H.-G. (2012). Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. *PloS One*, 7(12):e50781.

Xie, X., Lu, J., Kulbokas, E., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and $3'$ utrs by comparison of several mammals. *Nature*, 434(7031):338–345.

Yamanishi, Y., Vert, J.-P., Nakaya, A., and Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19(suppl 1):i323–i330.

Yang, P., Ho, J., Zomaya, A., and Zhou, B. (2010). A genetic ensemble approach for gene-gene interaction identification. *BMC Bioinformatics*, 11:524.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.*, 39(17):7415–7427.

Yuan, T. and Cantley, L. (2008). Pi3k pathway alterations in cancer: variations on a theme. *Oncogene*, 27(41):5497–5510.

Zhang, J., Zhan, Z.-h., Lin, Y., Chen, N., Gong, Y.-j., Zhong, J.-h., Chung, H. S., Li, Y., and Shi, Y.-h. (2011). Evolutionary computation meets machine learning: A survey. *IEEE Comput. Intell. Mag.*, 6(4):68–75.

Zhang, S., Liu, X., Zhang, Y., Cheng, Y., and Li, Y. (2013). Rnai screening identifies kat8 as a key molecule important for cancer cell survival. *Int. J. Clin. Exp. Pathology*, 6(5):870.

Zhong, H., McCord, R., and Vershon, A. K. (1999). Identification of target sites of the $\alpha$2–mcm1 repressor complex in the yeast genome. *Genome Res.*, 9(11):1040–1047.

Zhou, A., Zhang, Q., and Jin, Y. (2009). Approximating the set of pareto-optimal solutions in both the decision and objective spaces by an estimation of distribution algorithm. *IEEE Trans. Evol. Comput.*, 13(5):1167–1189.

Zhuang, J., Jones, A., Lee, S.-H., Ng, E., Fiegl, H., Zikan, M., Cibula, D., Sargent, A., Salvesen, H. B., Jacobs, I. J., et al. (2012). The dynamics and prognostic potential of dna methylation changes at stem cell gene loci in women's cancer. *PloS Genet.*, 8(2):e1002517.

Zuo, T., Tycko, B., Liu, T.-M., Lin, H.-J. L., and Huang, T. H. (2009). Methods in dna methylation profiling. *Epigenomics*, 1(2):331–345.

# 초 록

생명과학 연구의 기본적 목표 중 하나는 생물학적 인자들과 표현형의 복잡한 관계를 이해하고, 표현형에 영향을 미치는 다양한 인자들을 밝히는 것이다. 특히 유전체 서열은 유전자 발현이나 질병 민감도 등의 표현형을 결정하는 데에 있어서 중요한 역할을 한다. 따라서 유전체 서열 기반 정보에 대한 연구는 생물학적 기작을 이해하기 위해 필수적이다. 기존의 유전체 서열 관련 연구는 주로 생체 내 기작에 중요한 영향을 미치는 하나의 인자를 찾는 것에 집중되어 있었다. 최근 대용량 생물학 데이터 생산 기술의 발전으로 인해 전역 유전체 수준에서 유전적 변이를 분석하고 질병의 원인을 찾고자 하는 시도가 가능하게 되었지만, 거대한 탐색 공간과 계산 복잡도로 인해 여전히 다중 인자들의 고차 관계를 탐색하여 분석하는 것은 쉬운 일이 아니다.

본 논문에서는 진화 연산과 통계적 학습 방법을 결합하여 다중 인자 상호 작용을 탐색할 수 있는 효과적인 방법들을 제안한다. 본 논문의 방법들은 다양한 전역 유전체 서열 분석 문제에서 상호 연관된 인자 조합과 기능적 모듈의 탐색을 목적으로 한다. 우선 통계적 학습 방법을 이용하여 유전자 발현 조절에 함께 영향을 주는 서열 조각 및 DNA 메틸화 영역을 탐색한다. 이후 인간 유전체와 같이 많은 수의 인자들을 가진 고차원의 서열 데이터 분석을 위해 진화 연산 개념을 도입한다. 본 논문에서 사용된 방법은 학습 데이터를 이용한 기계 학습 기술을 기반으로 하여 진화 연산 과정에서 문제 공간을 효과적으로 탐색한다. 이를 통해 계산학적으로 복잡한 최적화 문제에서 답이 될 수 있는 후보군들을 찾아가는 것이 가능하다. 유전체 및 후성유전체 데이터를 이용한 실험 결과는 본 논문에서 사용된 진화 연산 기반 방법이 질병과 연관된 고차 상호 관계를 발견할 수 있다는 것을 보인다. 따라서 본 논문의 연구는 유전체 및 후성유전체 연구에서 서열 기반 인자들 간의 복잡한 상호작용을 분석할 수 있는 유용한 방법이 될 수 있을 것이다.

**Keywords:** **고차상호작용, 진화연산, 유전체 서열 분석, 기계학습, 유전체학, 후성유전체학**

**학번: 2004-20623**

理學博士學位論文

# Evolutionary Machine Learning of Higher Order Relationships in Genome-wide Sequence Analysis

유전체 서열 분석에서 고차 관계의 진화적 기계학습

2014年 2月

서울大學校 大學院

협동과정 생물정보학 전공

李 齊 根

# 유전체 서열 분석에서 고차 관계의 진화적 기계학습

# (Evolutionary Machine Learning of Higher Order Relationships in Genome-wide Sequence Analysis)

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 전공

李 齊 根

李齊根의 理學博士 學位論文을 認准함

2013年 12月

委 員 長 김 선

副委員長 장 병 탁

委　　員 김 희 발

委　　員 신 수 용

委　　員 백 대 현

# 유전체 서열 분석에서 고차 관계의 진화적 기계학습

# Evolutionary machine learning of higher order relationships in genome-wide sequence analysis

Je-Keun Rhee

Ph.D. Thesis

Interdisciplinary Program in Bioinformatics
Seoul National University

Feb., 2014

Supervisor: Byoung-Tak Zhang

# Abstract

One of the basic research goals in life science is to understand the complex relationships between biological factors and phenotypes, and to identify the various factors affecting the phenotype. In particular, genomic sequences play a significant role in determining the phenotype, such as gene expression and a susceptibility to disease, so the studies for the fundamental information stored in genome is essential to understanding biological processes. Previous genomic sequence analyses mainly focused on identification of a single associated factor or pairwise relationships with significant effects. Recent development of high-throughput technologies has made it possible to identify the causal factors by carrying out genome-wide analysis. However, it still remains as a challenge to discover higher-order interactions of multiple factors because this involves huge search spaces and computational costs.

In this dissertation, we develop effective methods for identifying the higher-order relationships of sequence elements affecting the phenotype, by combining statistical learning with evolutionary computation. The methods are applied to finding the associated combinatorial factors and dysfunctional modules in various genome-wide sequence analysis problems. Firstly, we show statistical learning-based methods to detect co-regulatory sequence motifs and to investigate combinatorial effects of DNA methylation, affecting on downstream gene expression. Next, to examine the sequence datasets with a huge number of attributes on human genome, we apply evolutionary computation approaches. Our methods search the problem feature space based on machine learning techniques using training datasets in evolutionary computation processes and are able to find candidate solution well in computationally expensive optimization problems. The experimental results show that the approaches are useful to find the higher-order relationships associated to disease using genomic and epigenomic datasets. In conclusion, our studies would provide practical methods to analyze complex interactions among sequence elements in genomic/epigenomic studies.

**Keywords: Higher-order interaction, Evolutionary computation, Genome-wide sequence analysis, Machine learning, Genomics, Epigenomics**

**Student Number: 2004-20623**

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

The post-genomic era is characterized by a tremendous revolutionary expansion in biological data. Over past few decades, there has been rapid development in biological research and technologies, as a result, a huge amount of data have been produced. In particular, with the advances of sequencing technologies, a large amount of datasets have been deposited in repositories (Metzker, 2009). Understanding and exploiting these data is now a key to success of advancing biological research, and the requirements have stimulated development and expansion of applying computational approaches in biology.

The large expansion of genome-wide measurement data poses the research question of how to retrieve the valuable knowledge from the genomic sequences (Huttenhower and Hofmann, 2010; Chin et al., 2011). Traditionally, genomic studies mainly focused on central dogma in molecular biology, namely from genome to transcriptome. Experimentally determined catalogues of genes only tell us about a basic building block of the biological regulatory processes. They do not tell us much about how the biological processes operate as a system, such as higher order functional behaviors (Chuang et al., 2010). Although many computational approaches have dealt with high-throughput biological datasets generated in multi-dimensional forms, it is still important to search the large datasets efficiently and effectively (Palsson and Zengler, 2010; Kouskoumvekaki et al., 2013).

Actually, most biological problems are complex and hard to be understood. One problem is to investigate the interactions of the various factors, since the biological processes are affected by multiple factors. Although genome-wide analysis can be possible with the development of high-throughput technologies, an exhaustive search of all potential solutions is still challenging, and most likely impossible. The standard constructive and approximate approaches are usually impractical in terms of a huge search space and lots of computational costs. Thus, the genome-wide sequence analyses mainly focused on identification of a single associated factor or pairwise relationships with significant main effects (Cordell, 2009; Bush and Moore, 2012).

The genome-wide sequence analyses have contributed to ability to identify genomic sequence elements that are associated with phenotypes such as gene expression and disease (Feero et al., 2010; Heap et al., 2009; Kim et al., 2012; Kang et al., 2011). It has been possible to find a single sequence element that has statistically significant association with phenotype. To date, many associated genes or sequence elements were found, but there were not clear explain the complex biological processes (McCarthy et al., 2008; Stranger et al., 2011). Thus, there has been increased interest in discovering combinations of the sequence elements that are strongly associated with a phenotype even if each element has little or even no individual effect. Thus, an alternative research view of post-genomic/epigenomic era would be to go up eventually to still higher levels, i. e. biological systems.

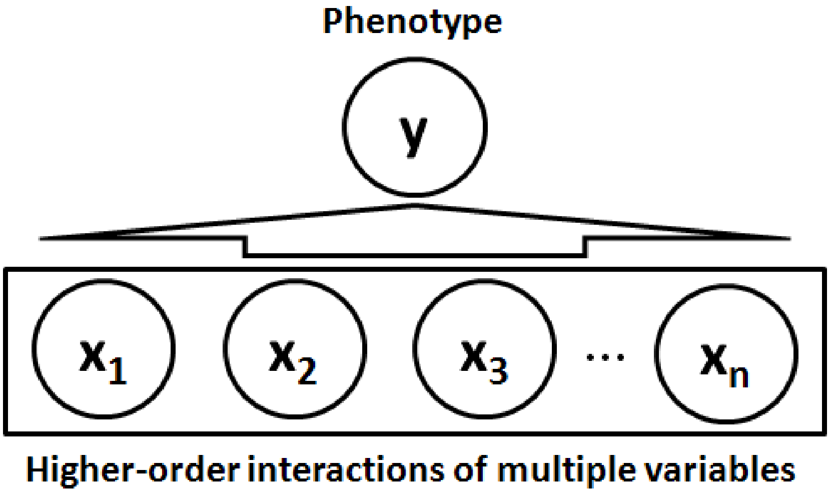


Figure 1.1: Schematic concept for higher-order interaction and its effects on phenotype

## 1.2 Approaches

In this dissertation, we present computational approaches to predict higher-order relationships of disease genes or sequence elements and identify dysfunctional modules, based on machine learning and evolutionary computation using phenotype and sequence information. Our goal is to discover and study the combinations of sequence elements affecting on phenotype. In particular, we focus on discovering the interactions, especially high-order ones beyond size 2, that are strongly associated with a phenotype and yield information on interpretable statistical and functional interactions (Figure 1.1).

At first, we show a way to search co-regulatory sequence motifs using a statistical learning method, kernel canonical correlation analysis (kernel CCA) (Rhee et al., 2009). One of the major challenges in gene regulation studies is to identify regulators affecting the expression of their target genes in specific biological processes. Despite their importance, regulators involved in diverse biological processes still remain largely unrevealed. In the study, we propose a kernel-based approach to efficiently identify core regulatory elements and their combinations involved in specific biological processes using gene expression profiles. We develop a framework that can detect correlations between gene expression profiles and the upstream sequences on the basis of the kernel canonical correlation analysis (kernel CCA). We show that upstream sequence patterns are closely related to gene expression profiles based on the canonical correlation scores obtained by measuring the correlation between them. The experimental results show that our method is able to successfully identify regulatory motifs and their co-regulatory pairs involved in specific biological processes.

Secondly, we investigated the combinatorial effects of DNA methylation on downstream gene expression using machine learning approaches (Rhee et al., 2013). Aberrant DNA methylation of CpG islands (CGIs), CGI shores, and first exons is known to play a key role in the altered gene expression patterns in all human cancers. To date, a systematic study on the effect of DNA methylation on gene expression using high resolution data has not been reported. In this study, we conducted an integrated analysis of MethylCap-sequencing data and Affymetrix gene expression microarray data for 30 breast cancer cell lines representing different breast tumor phenotypes. We develop methylome data analysis protocols for the integrated analysis of DNA methylation and gene expression data on the genome scale and we present comprehensive genome-wide methylome analysis results for differentially methylated regions and their potential effect on gene expression in 30 breast cancer cell lines representing three molecular phenotypes, luminal, basal A, and basal B. Our integrated analysis demonstrates that methylation status of different genomic regions may play a key role in establishing transcriptional patterns in molecular subtypes of human breast cancer.

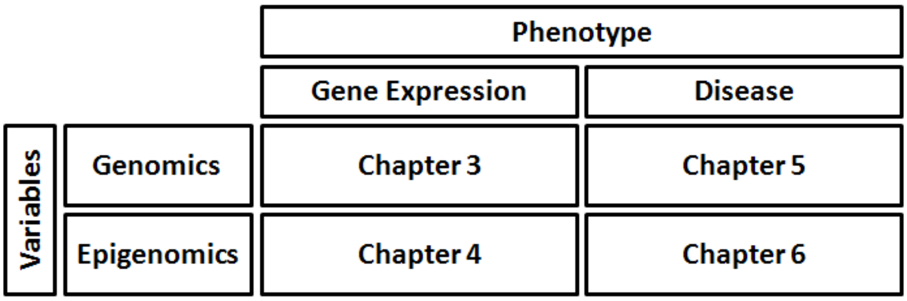
These two genome-wide approaches were useful for identification of co-regulatory interactions or combinatorial effects, associated to downstream gene expression. However, sometimes, it might need another approach to examine a huge number of sites on whole genome and to discover higher-order relationships of sequence elements associated with complex disease. Then, we applied evolutionary computation approaches to identify higher-order interaction of multiple factors associated to disease. Evolutionary computation is a general purpose search approach that uses principles inspired by natural genetic populations to evolve solutions to problems (Simon, 2013). The basic idea is to maintain a population of individuals which represent plausible solutions to the problem, which evolves over time through a process of competition and controlled variations.

In the framework of evolutionary machine learning, the main idea is that the evolutionary computation method has stored training data to search problem feature space and population information during the iterative evolutionary process. Then, the machine learning technique is helpful in analyzing these data for enhancing the search performance.

We propose an approach to search the higher-order interaction for genome-wide association studies based on the evolutionary machine learning. Searching for the relationship between the genetic variant and its phenotypic effects is important to understand the genetic basis and mechanism of many complex genetic diseases. There have been a lot of research to analyze the causality and, in many studies, it have led to succeed to discover the associations of genes with diseases. Although there exist lots of the genetic variants with major effects and they can be linked to complex diseases, however, it is still challenging to find the multiple interactions from a millions of SNPs and their association with a disease. Here, we present an approach to analyze higher-order interactions of the genetic variations, which associated with a disease. The method searches combinatorial feature spaces of the genetic variants and selects the higher-order variables which are distinctive to classify the disease and normal samples by evolutionary learning. We test the method and illustrate the advantages with genetic variant datasets for type 2 diabetes. As a result, our approach could identify the higher-order interaction of SNPs associated with type 2 diabetes, and especially detect several interactions specific in Korean population.

Finally we introduced probabilistic concepts in the evolutionary computation for identification of DNA methylation modules. By exploring the problem space by building and sampling explicit from probabilistic graphical models, the approach would be proper to find the higher-order relationships or biological modules.

Considerable studies have been made to elucidate effects of genetic variability in complex disease, but it is still challenging to discover molecular pathogenesis clearly. The epigenetic factor would be another candidate to make up the complex regulatory mechanism. Especially it is well-known that DNA methylation could lead to inhibition of downstream gene expression. Although many researchers are trying to clarify the relationships between DNA methylation and gene expression, recently, more efforts are required to find the multiple interactions from a lot of DNA methylation sites and their association with a disease. To assess DNA methylation modules potentially relevant to disease, we use an estimation of distribution algorithm (EDA)-based learning method identifying high-order interaction of DNA methylation sites. It finds a solution which is a set of discriminative methylation sites by building a probabilistic dependency model. The algorithm is applied to array- and sequencing-based high-throughput DNA methylation profiling datasets, and the experimental



| | | Phenotype | |
|---|---|---|---|
| | | Gene Expression | Disease |
| Variables | Genomics | Chapter 3 | Chapter 5 |
| | Epigenomics | Chapter 4 | Chapter 6 |

Figure 1.2: Organization of chapters

results show that it has a good search ability to identify the DNA methylation modules for a specific disease.

Our approaches would provide practical methods to integrate large amount of datasets and to analyze complex interactions among building blocks and with dynamic environments.

## 1.3 Organization of the dissertation

This dissertation is organized as follows (Figure 1.2):

- In Chapter 2, we briefly introduces informatics and computational approaches in genomic analysis. We describe background of genome biology, and explain what the machine learning and evolutionary computation are. Then, the basic concepts and their several applications in biological domains are described.

- In Chapter 3, we search co-regulatory sequence motifs by a kernel-based correlation analysis. We identify regulatory sequences affecting the expression of their downstream genes. And we investigate pairwise relationships of the sequence motifs closely related to gene expression profiles in a specific biological process.

- Chapter 4 discribes analysis protocols to investigate effects of DNA methylation in various sites on downstream gene expression. Using high resolution sequencing-based methylation profiling datasets, we show comprehensive genome-wide methylome analysis results for their potential effect on gene expression. The analysis results present that methylation status of different genomic regions may play combinatorial effects on transcriptional patterns via a statistical learning approach.

- In Chapter 5, we propose an evolutionary learning method for identifying higher-order interaction of multiple SNPs in genome-wide association studies. We show that the proposed evolutionary learning method searches combinatorial feature spaces and identifies the higher-order variables which are related to disease.

- In Chapter 6, we use a probabilistic evolutionary learning to find higher-order relationships from a lot of DNA methylation sites, which is potentially relevant to disease. Instead of crossover or mutation operators in traditional evolutionary computation, we build a probabilistic distribution model and are sampled from the model in the evolutionary learning processes. The experimental method and results represent that the approach can be a new systematic way to identifying high-order interaction of DNA methylation sites and DNA methylation modules which is associated to disease.

- Finally, we summarize the dissertation and discuss our research in Chapter 7.

# Chapter 2

# Genome biology and computational analysis

## 2.1 Fundamentals of genome biology

### 2.1.1 DNA, gene, chromosomes and cell biology

DNA (deoxyribonucleic acid) is a biomolecule that includes information for how organisms are genetically built. DNA is a double strand structure that contains complementary genetic information encoded by 4 bases, adenine (A), guanine (G), thymine (T) and cytosine (C). A gene is a segment of DNA that can be inherited from parents to children and can confer a trait to the offspring. The genes are organized and packaged in chromosomes. In case of human, there exist 23 pairs of chromosomes.

One set of chromosomes for each pair comes from a person's mother, and the other set is from father. New cells get their chromosomes from old cells through cell division, mitosis. The chromosome in cell nucleus is divided into two identical sets by mitosis of cell cycle. The primary result of mitosis is the transferring of the parent cell's genome into two daughter cells. Cell cycle is the series of events leading to its growth, replication (duplication) and division of a eukaryotic cell. The cell cycle can be divided into several phases: G1, S, G2 and M phases. At G1 and G2 phases, cells increase in size and DNA replication occurs at S phase. M phase is a periods of mitosis which is cell division state. The cell growth stops at this stage and the cell divides itself into two distinct daughter cells.

### 2.1.2 Gene expression and regulation

Gene expression is a fundamental step at which a genotype gives rise to a phenotype. The gene expression means a process that the genetic information from a gene is used in production of a functional gene product (protein or RNA). The process is generally described by that a gene is transcribed into RNA and this transcript may then be translated into protein.

Regulation of gene expression includes mechanisms to increase or decrease the production of specific gene products. The program of gene expression is very sophisticate. A complex set of interactions between genes, RNA molecules, proteins (including transcription factors) and other components of the expression system determine when and where specific genes are activated and the amount of protein or RNA product produced. Some genes are expressed continuously, as they produce proteins involved in basic metabolic functions; some genes are expressed as part of the process of cell differentiation; and some genes are expressed as a result of cell differentiation.

Specific DNA sequences are accessible for specific proteins to bind. Many of these proteins are activators, while others are repressors. Such proteins are often called transcription factors (TFs). Transcription factors are proteins that play a role in regulating the transcription of genes by binding to specific regulatory nucleotide sequences. Each TF has a specific DNA binding domain that recognizes a 6-10 basepair motif in the DNA, as well as an effector domain (Matys et al., 2003; Sandelin et al., 2004).

For an activating TF, the effector domain recruits RNA polymerase II, the eukaryotic mRNA-producing polymerase, to begin transcription of the corresponding gene. TFs bind at the promoters just upstream of eukaryotic genes. However, they also bind at regions called enhancers, which can be oriented forward or backwards and located upstream or downstream or even in the introns of a gene, and still activate or repress the gene expression. Studying gene expression across the whole genome via microarrays or massively parallel sequencing allows investigators to see which groups of genes are co-regulated during differentiation, cancer, and other states and processes.

### 2.1.3 Genomics

Genome is the entirety of all genes and information contained within the noncoding regions from an organism, mainly encoded by DNA. Genomics usually describe studies to determine the entire DNA sequence of organisms and genomic structures. The field also includes studies of various genomic phenomena. In contrast to the classical molecular biology or genetics to investigate the roles and functions of single gene, genomics aim to elucidate its effects on the entire genomic networks with its genetic and functional information (Lander, 1996).

A major branch of genomics is concerned with sequencing the genomes of various organisms. A rough draft of the human genome was completed in 2001 (Venter et al., 2001; Lander et al., 2001). Since then, there have been much more studies for human genome. Also, the genomic information of many other species has been successfully achieved. The knowledge of full genomes has created the possibility for the field of functional genomics, mainly concerned with patterns of gene expression during various conditions. For the purpose, computational approaches would be the most important tools here.

### 2.1.4 Epigenomics

The classical biology states that DNA is transcribed to RNA, RNA is translated to protein, and it regulates various cellular processes and functions. In the traditional views, phenotypic alteration has been caused by aberrant sequence variants or an inherited genomic allele. However, in the recent view, cells with identical DNA sequences can have a variety of distinct functions and phenotypes, by epigenetic modification including DNA methylation and histone modification (Murrell et al., 2005; Holliday, 2006). That is, the epigenetic modifications affect gene expression without altering the DNA sequences and play an important role in numerous cellular processes such as in differentiation, development and tumorigenesis (Bernstein et al., 2007; Baylin and Jones, 2011).

One of the most characterized epigenetic modifications is DNA methylation. DNA methylation is a process by which a methyl group is added to DNA. The methylation is most commonly found on cytosine residues adjacent to guanine, termed CpG dinucleotides (Laird, 2010). It is well-known that the DNA methylation can control gene expression. Usually the DNA methylation represses gene expression by a multistep process, although the exact mechanim is unknown.

Epigenomic research tries to identify and characterize epigenetic modifications on a global level. The study of epigenetics on a global level has been made possible recently through high-throughput assays. To manage a huge size of datasets and to clarify the complex mechanism on the fields, as in the other genomics fields, epigenomics also relies heavily on bioinformatics, which combines the disciplines of biology, mathematics and computer science.

## 2.2 Evolutionary machine learning

### 2.2.1 Machine learning and evolutionary computation

Machine learning is a study to give computers abilities to learn from existing data. Usually, it can be used to discover patterns and rules from data, and predict future events. Machine learning techniques generally involves statistical methods, interpolation and regression, supervised classification algorithms, clustering analysis, reinforcement learning, and so on.

The ideas and techniques from machine learning can be hybridized with evolutionary computation. Evolutionary computation with machine learning techniques would be a promising research direction to search optimal solution from the machine learning point of view (Zhang et al., 2011). Evolutionary computation is a kind of optimization methodology inspired by mechanisms of biological evolution. It can be widely used as an optimization tool in recent years.

The first step of the evolutionary computation is initialization of population. Next, it enters iterative evolutionary step with fitness evaluation, selection, and population reproduction. The newly generated population is evaluated again and the iteration continues until a termination criterion is satisfied.

### 2.2.2 Evolutionary computation in biology

The genomic revolution is generating a huge amount of data in rapid speed but it has become made difficult for biologists to decipher. In addition, many problems in biology are too large to solve with standard methods. Evolutionary computation can be a solution for the current bioinformatics problems (Fogel and Corne, 2002;

Pal et al., 2006). Although bioinformatics present a number of difficult optimization problems, evolutionary computation can rapidly search very large and complex spaces and return reasonable solutions.

The evolutionary computation has experienced a large growth in applications for bioinformatics with several advantages. For example, the errors generated in biological experiment data might be handled with no significant problem in the evolutionary computation. The errors can contribute to genetic diversity, a desirable property in the evolutionary learning processes. Thus, it might be more tolerable in using evolutionary computation than other deterministic algorithms. Sometimes, several tasks of bioinformatic studies do not require the exact optimum answer. Instead, they require robust and close approximate solutions. Also, local optimal solution can be helpful to understand biological processes. Evolutionary computation approaches can be also efficient to provide the solution in this case. In addition, EAs can process, in parallel, population billions times larger than is usual expectation is that larger populations can sustain larger range of genetic variation, and thus can generate high-fitness individuals in fewer generation. Laboratory operations on DNA inherently involve errors. These are more tolerable in executing evolutionary algorithms than executing deterministic algorithms.

Evolutionary computation has been profitably used in traditional bioinformatic problems. Several application areas follow:

- Sequence alignments

  Multiple sequence alignment helps to infer evolutionary history or discover conserved regions among closely related sequences. The problem is known as NP-hard. Genetic algorithms can be used to find optimal solutions in this problem (Notredame and Higgins, 1996; Nguyen et al., 2002).

- Motif finding

An instance of genetic algorithms can be used for motif finding, similar to Gibbs sampling. The motifs can generated from randomly selected sequences, and then alignment scores has been computed between the sequence fragments and the motifs. It increases the chance to find the real sequence motifs (Liu et al., 2004; Das and Dai, 2007).

- Protein structure prediction

  Evolutionary computation methods for protein structure prediction have been developed in the last decades. These have attempted to optimize the energy function of the peptide chain and to determine the optimal protein folding (Unger and Moult, 1993; Cooper et al., 2003).

- Protein-protein interaction and docking

  Protein interaction and docking represents fundamental function of biomolecules. Although it is possible now to determined by experimental methods, it is difficult to predict the recognition exactly ascertaining the structure of protein complexes. The evolutionary computation approaches can help to solve the problem (Morris et al., 1998; Wang et al., 2010).

The applications suggest that a variety of problems in biological domains can be well-suited for evolutionary computation approaches and be analyzed well by the methods.

# Chapter 3

# Identifying co-regulatory sequence motifs

## 3.1 Background

One of the major challenges in current biology is to elucidate the mechanism governing the gene expression. Gene expression programs depend mainly on transcription factors which bind to upstream sequences by recognizing short DNA motifs called transcription factor binding sites (TFBSs) to regulate their target gene expression (Lee et al., 2002). Transcription factors bind to upstream sequences to regulate gene expression. They recognize short DNA motifs called transcription factor binding sites (TFBSs). Although many regulatory motifs have been identified, large amount of functional elements still remain unknown (Xie et al., 2005).

Many genome-wide approaches have been developed in attempt to discover regulatory motifs from upstream sequences. The early computational approach for identifying regulatory motifs is based on statistical analyses using only upstream sequences of genes. Statistical methods such as maximum-likelihood estimation or Gibbs sampling, are effective for searching directly significant sequence motifs from multiple upstream sequences (Hughes et al., 2000; Bailey and Elkan, 1994). Several computational approaches based on machine learning methods have also been implemented. A SOM (self-organizing map)-based clustering method can find regulatory sequence motifs by grouping relevant sequence patterns (Mahony et al., 2005) and a graph-theoretic approach has tried to identify regulatory motifs by searching the maximum density subgraph (Fratkin et al., 2006).

More advanced approaches have been developed that can identify regulatory motifs by linking gene expression profiles and motif patterns. The main advantage of these approaches is that they can identify motifs correlated to specific biological processes. Most early trials used a unidirectional search, such as approaches that search for shared patterns with upstream sequences in a set of co-expressed genes that were found by clustering algorithms (Tavazoie et al., 1999; Brāzma et al., 1998) or those that determine whether genes with common regulatory elements are co-expressed (Pilpel et al., 2001; Park et al., 2002). In addition, it is also possible to link motifs to gene expression patterns using linear regression models or regression trees (Bussemaker et al., 2001; Keles et al., 2002). Recently, several techniques for a bidirectional search to detect the relationship between the regulatory motifs and the gene expression profiles have been emerged (Segal et al., 2003; Jeffery et al., 2007). They search regulatory motifs more efficiently than unidirectional approaches since they search similar expression patterns and regulatory motifs correlated to them simultaneously.

In this study, we propose a novel bidirectional approach using a kernel-based method, kernel CCA (kernel canonical correlation analysis), to analyze the relationship between regulatory sequences and gene expression profiles (Hardoon et al., 2004; Akaho, 2006; Bach and Jordan, 2003). The expression and sequence features are mapped from the original input space to a higher dimension space using a kernel trick, and the relationship between the two projected objects is interpreted to identify highly correlated motifs (Figure 3.1). Our method has advantages that it can detect core motifs relevant to a specific cellular process without the additional efforts of clustering and intensive motif sampling process in upstream sequences.

We applied the kernel CCA to a paired set of upstream sequence motifs of genes and their expression profiles in yeast *Saccharomyces cerevisiae* cell cycle, and explored significant relationships between motifs and expression profiles. We also searched for regulatory motifs correlated with specific expression patterns. We also searched for regulatory motifs correlated with specific expression patterns. Our method retrieved regulatory motifs that play an important role in cell cycle regulation including several well-known cell cycle regulatory motifs: MCB, SCB and SFF'. Furthermore, we identified motif pairs associated with the gene expression to construct a map of combinatorial regulation of regulators.

## 3.2 Methods

### 3.2.1 Investigation of the relationship between regulatory sequence motifs and expression profiles

Kernel CCA (Canonical correlation analysis) is a version of the nonlinear CCA, where the kernel trick is utilized to find nonlinearly correlated features from two datasets (Hardoon et al., 2004; Akaho, 2006; Bach and Jordan, 2003). Canonical correlation analysis (CCA) CCA is a classical multivariate statistical method for finding linearly correlated features from a pair of datasets (Hotelling, 1936). Suppose there is a pair of multivariates $\mathbf{x}$ and $\mathbf{y}$, CCA finds a pair of linear transformations such that the correlation coefficient between extracted features is maximized. How-

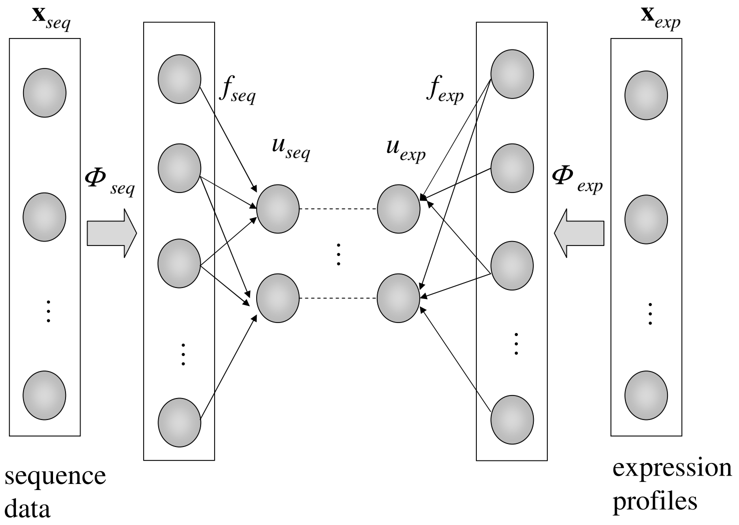


Figure 3.1: The basic scheme of the kernel CCA. The sequence and expression data are transformed to Hilbert space by $\phi$ function. By taking inner products, $u_{exp}$ and $u_{seq}$ were derived, which maximize the correlation between the upstream sequences and the expression profiles.

ever, if there is a nonlinear relationship between the variates, CCA does not always extract useful features.

Kernel CCA offers a solution for overcoming the linearity by first projecting the data into a higher dimensional feature space. While CCA is limited to linear features, kernel CCA can capture nonlinear relationships. Kernel CCA has been used for several applications including text retrieval and biological data analysis (Hardoon et al., 2004; Yamanishi et al., 2003).

Figure 3.1 illustrates the basic scheme of the kernel CCA for our integrated analysis of DNA sequence motif and gene expression data. Using kernel CCA, we tried to find maximally correlated features between the gene expression and the sequence motifs. Here, a gene set $\mathbf{X}$ is represented by two separate profiles in terms of its transcriptional behavior and upstream sequences, $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$. These are composed of the expression profile, $\mathbf{x}_{exp} = (e_1, e_2, ..., e_N)$ and the sequence profile, $\mathbf{x}_{seq} = (m_1, m_2, ..., m_M)$ of each gene. Here $e_i$ $(1 \leq i \leq N)$ is the expression value of the gene in the $i$-th sample or experimental condition from microarray data, and $m_j$ $(1 \leq j \leq M)$ denotes the occurrence frequency of the $j$-th sequence motif in the upstream region of the gene. For the detection of the correlated features between the two datasets, $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$ are first mapped to Hilbert space, $H$, by function $\phi$. That is, each $\mathbf{x}$ is projected into two directions, $f_{exp}$ and $f_{seq}$, in Hilbert space according to its representation:

$$u_{exp} = \Big\langle f_{exp}, \phi_{exp}(\mathbf{x}_{exp}) \Big\rangle \tag{3.1}$$

$$u_{seq} = \Big\langle f_{seq}, \phi_{seq}(\mathbf{x}_{seq}) \Big\rangle, \tag{3.2}$$

where $\Big\langle \cdot, \cdot \Big\rangle$ denotes the dot product. Kernel CCA looks for maximally correlated features between $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$:

$$\begin{aligned}&\gamma(f_{exp}, f_{seq}) = \\ &\quad \max \frac{\mathbf{cov}(u_{exp}, u_{seq})}{(\mathbf{var}(u_{exp}) + \lambda_{exp}\|f_{exp}\|^2)^{\frac{1}{2}} (\mathbf{var}(u_{seq}) + \lambda_{seq}\|f_{seq}\|^2)^{\frac{1}{2}}},\end{aligned} \tag{3.3}$$

where $\lambda_{exp}$ and $\lambda_{seq}$ are regularization parameters. The kernel CCA can be given by solving a generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \mathbf{K}_{exp}\mathbf{K}_{seq} \\ \mathbf{K}_{seq}\mathbf{K}_{exp} & 0 \end{pmatrix} \begin{pmatrix} \alpha_{exp} \\ \alpha_{seq} \end{pmatrix} =$$

$$\rho \begin{pmatrix} (\mathbf{K}_{exp} + \frac{n\lambda_{exp}}{2}\mathbf{I})^2 & 0 \\ 0 & (\mathbf{K}_{seq} + \frac{n\lambda_{seq}}{2}\mathbf{I})^2 \end{pmatrix} \begin{pmatrix} \alpha_{exp} \\ \alpha_{seq} \end{pmatrix}, \tag{3.4}$$

where $\mathbf{I}$ denotes the identity matrix, $\mathbf{K}_{exp}$ is the kernel matrix for expression profile data, and $\mathbf{K}_{seq}$ is the kernel matrix for sequence motif data. When given $\alpha_{exp}$ and $\alpha_{seq}$ as the solution of the above generalized eigenvalue problem with the largest eigenvalue, canonical correlation scores (CC scores) for $\mathbf{x}_{seq}$ and $\mathbf{x}_{seq}$ are estimated by $u_{seq} = \mathbf{K}_{seq}\alpha_{seq}$ and $u_{exp} = \mathbf{K}_{exp}\alpha_{exp}$. The CC scores are the low dimensional mapping of genes in terms of two separate representations and can be used to show the salient correlation between the two. Once we have obtained the $\alpha$ vector, the weights of the motif and expression profile, $\mathbf{W}_{seq}$ and $\mathbf{W}_{exp}$, are obtained as follows:

$$\mathbf{W}_{exp} = \mathbf{x}_{exp}^T\alpha_{exp} \tag{3.5}$$

$$\mathbf{W}_{seq} = \mathbf{x}_{seq}^T\alpha_{seq}. \tag{3.6}$$

A high weight value of the specific sequence motif means that the motif is strongly correlated with the expression patterns of genes whose upstream region includes the motif and whose CC scores are high. If a weight of a specific motif has a high absolute value, the motif is more likely to be investigated further.

### 3.2.2 Preparation of the gene expression datasets

Expression profiles of all ORFs (open reading frames) during the yeast cell cycle that consists of 18 time points in the alpha factor synchronization case [18] were used as the expression dataset. To map from the expression profiles to high dimensional

Table 3.1: Known regulatory motifs in yeast *Saccharomyces cerevisiae.*

| Motif Name | | | |
|---|---|---|---|
| RAP1 | RPN4 | GCN4 | MCB |
| HAP234 | MIG1 | AFT1 | STRE' |
| CCA | CSRE | PHO4 | STE12 |
| HSE | ABF1 | ATRepeat | GAL |
| Leu3 | LYS14 | MET31-32 | OAF1 |
| PAC | PDR | PHO | REB1 |
| STRE | ECB | ndt80(MSE) | Yap1 |
| SCB | Gcr1 | zap1 | MCM1' |
| MCM1 | SFF | SFF' | BAS1 |
| Ume6(URS1) | SWI5 | ALPHA1' | ALPHA1 |
| ALPHA2' | ALPHA2 | | |

space, we converted them to the kernel matrix. We applied a gaussian RBF kernel to the expression profile matrix by: $k(\mathbf{x}_{exp}, \mathbf{x'}_{exp}) = \exp[-\frac{d(\mathbf{x}_{exp},\mathbf{x'}_{exp})}{2\sigma^2}]$, where $\sigma$ is a parameter and function $d(\cdot,\cdot)$ is a Euclidean distance.

### 3.2.3 Preparation of the gene sequence datasets

The sequence data was used in two ways. In the first case, we used the sequences of a total of 42 known motifs (Table 3.1) extracted by Pilpel (Pilpel et al., 2001). It was composed of 42 motifs (Table 1). We then scanned the upstream regions of ORFs for the presence of these motifs using the AlignACE program (Hughes et al., 2000). The sequence profile was represented by the occurrence of these motifs in the promoters of each gene in the genome.

In the second case, we analyzed the relationship between the expression profiles and the raw upstream sequences. We extracted gene upstream sequences $\sim$ 1kb from each gene. From these sequences, we calculated the frequency of all possible $l$-mers in each gene. For $l = 5$, each gene had 1024 ($= 4^5$) base combinations. The sequence profile was encoded in the frequency of $l$-mers.

We applied the kernel as $k(\mathbf{x}_{seq}, \mathbf{x}'_{seq}) = (\mathbf{x}_{seq}^T \mathbf{x}'_{seq})^d$ to the sequence data. When $d = 1$, it is the linear kernel, and when $d > 1$, it is the polynomial kernel.

### 3.2.4 Measurement of the effect of motif combinations

To measure the effect of the motif pairs, we defined the ECRScore (Expression Coherence coRrelation Score) calculated by a Pearson correlation coefficient of expression profiles for all possible pairs of genes whose upstream regions had the two motifs, $m_i$ and $m_j$:

$$ECRScore(m_i, m_j) = \frac{N_\tau(m_i \bigcap m_j)}{N(m_i \bigcap m_j)}, \tag{3.7}$$

where $N(m_i \bigcap m_j)$ is the number of all pairs of genes whose upstream regions have the two motifs, and $N_\tau(m_i \bigcap m_j)$ is the number of gene pairs whose correlation coefficient is larger than the threshold $\tau$. The threshold was chosen based on the fifth percentile of the distribution for correlation coefficients of randomly sampled gene pairs.

## 3.3 Results

We applied a computational method, kernel CCA, to the identification of novel transcriptional regulatory elements. The main purpose of our experiments was to find regulatory motifs that were associated with gene regulation in specific biological processes. Using the kernel CCA, we first found highly correlated features between expression profiles and the sequence motifs. The key motifs in gene regulation were then identified from the weight scheme by the kernel CCA (see Methods section). Furthermore we demonstrate that it is possible for our method to be applied for identification of motif pairs using raw upstream sequences.

### 3.3.1 Identification of the relationship between gene expression and known motifs

We first explored the relationship between gene expression profiles and known motifs using a yeast gene expression dataset related to the cell cycle (Spellman et al., 1998) and a set of known motifs (see Table 3.1) extracted by AlignACE (Pilpel et al., 2001). A total of 551 ORFs (open reading frames) in the expression dataset contained at least one known motif. In the parameter setting, the degree of polynomial kernel was set to 3, the parameter $\sigma$ in Gaussian RBF kernel was 0.5, and the regularization parameter was 0.1. These parameters were chosen based on the parameter setting that produced a high correlation from multiple runs.

The results from the kernel CCA were visualized using the CC1 (first canonical correlation) score (Figure 3.2). In Figure 3.2, each point corresponds to a gene, and a cloud of the diagonal points illustrated the correlation between the expression and the motifs. The shape of diagonal points and the high correlation coefficient (0.996) indicated that the kernel CCA was able to find the close relationship between the expression profiles and the sequence motifs. We then performed the linear canonical correlation analysis using the same datasets. The correlation coefficient (0.612) obtained from the linear CCA was much lower. As shown in Figure 3.3, the linear CCA could not identify the significant correlation between expression profiles and motifs. This further supports that kernel CCA improve significantly in finding the correlation between the two datasets.

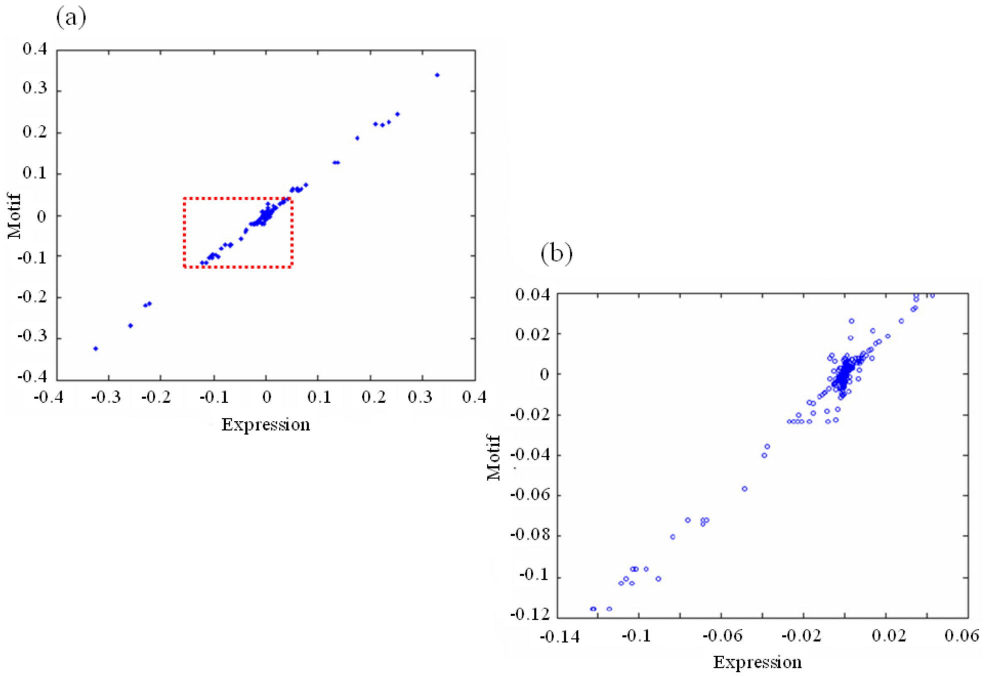


Figure 3.2: Relationship between gene expression profiles and regulatory sequence motifs. (a) The plot shows the correlation between gene expression profiles and the regulatory sequence motifs. Each dot represents one gene in the dataset, and x-axis means the value of $u_{exp}$, y-axis is $u_{seq}$. (b) The plot is a close-up view of the boxed area in (a).

The motifs were searched by the weight function of Equation 3.6 (see Methods section) with the model obtained by the kernel CCA and the top ranked motifs are shown in 3.2. SWI5 motif, a binding site of SWI5 protein, has the highest weight value. SWI5 has been known to act in G1 phase and in the M/G1 boundary in the cell cycle (Dohrmann et al., 1992, 1996). SFF' motif is a binding site of FKH1

Table 3.2: The list of top ranked motifs based on the weight scheme by the kernel CCA.

| Motif | Weight | Function |
|---|---|---|
| SWI5 | 0.89026 | Transcription Activation in G1 phase |
| SFF' | 0.45399 | FKH1 binding site that regulate the cell cycle |
| MCB | 0.29633 | MBF binding site that activates in late G1 phase |
| LYS14 | 0.21796 | Lysine biosysthesis pathway |
| ALPHA2 | 0.16532 | Encoding a homeobox-domain |

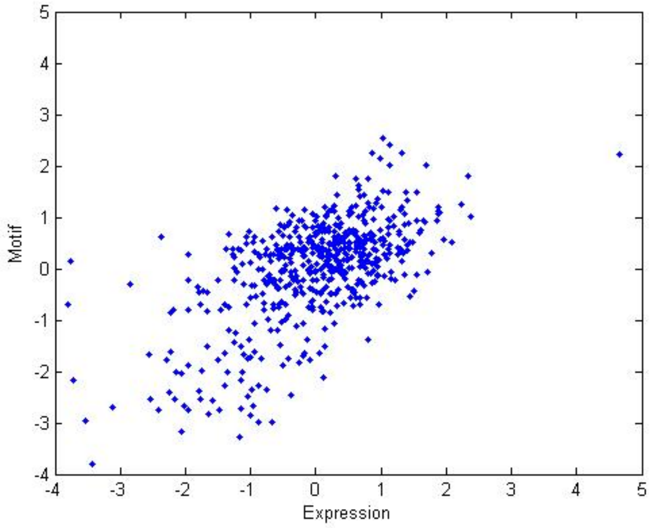


Figure 3.3: Relationship between gene expression profiles and regulatory motifs from the linear CCA

transcription factor that affects the expression of genes controlling the cell cycle during the G2-S phase change (Morillon et al., 2003). The MCB motif is one of the well-known motifs in the yeast cell cycle as a binding site in the MBF protein complex. MBF protein is composed of MBP1 and SWI6, and MBP1 is a DNA binding component while SWI6 has regulatory roles. It is well known that the MBF protein complex regulates the transcription of many genes in the late G1 phase (Dohrmann et al., 1992; Simon et al., 2001). ALPHA2 protein also plays a role in the cell cycle. It operates synergistically with MCM1 protein to repress the expression of its target genes (Vershon and Johnson, 1993; Zhong et al., 1999). MCM1 protein is a key regulator involved in the transcription of several M/G1 genes during the cell cycle (Simon et al., 2001; Lydall et al., 1991). A high weight value of ALPHA2 is supported by the evidence that ALPHA2 protein binds to the MCM1 protein and influences the regulation of other cell cycle-related genes (Keleher et al., 1989; Mead et al., 1996). Using the set of known motifs, our results are consistent with previous reports, validating the analysis method employed.

To further validate the result of top-ranked motifs extracted by kernel CCA, we compared the weights obtained from cell cycle-related ORF set with those obtained from randomly selected set. We performed the same procedure using random ORFs that are not known to be related to the cell cycle. Figure 3.4 shows the highly weighted motifs obtained from our method in cell cycle-related gene set and non cell cycle set, and the relative positions of those motifs are presented in the weight distribution of all motifs. The weight values obtained from random set were significantly lower than those obtained from cell cycle-related ORF set. We could infer that the significantly correlated motifs were not extracted from these random datasets. In summary, our method could identify the regulatory motifs that have high weights indicating high correlation between the upstream sequences and the gene expression profiles.

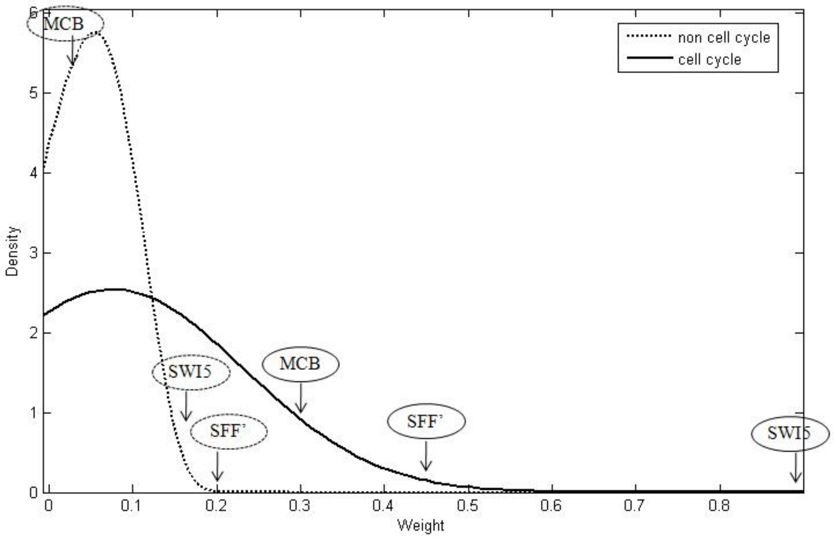


Figure 3.4: Weight distributions for MCB, SFF' and SWI5 motifs derived from cell cycle and non cell cycle-related datasets The dotted line indicates the weight distribution from the non-cell cycle datasets and the solid line from cell cycle datasets.

### 3.3.2 Identification of cell cycle-related motifs

We then applied the linear kernel to the motif sequence data containing a total of 1,024 features (window size $l = 5$) extracted from the raw upstream sequences of genes and Gaussian RBF kernels with parameter $\sigma$ values of 0.3 to the expression data. The regularization parameter was set to 0.1. These parameters are also empirically chosen based on the fact that they produced a high correlation. Figure 3.5 shows the CC1 score which represents the correlation between the expression profiles and the sequence patterns. When the linear kernel was applied to the sequence dataset, the expression data is closely related to the motif data using the

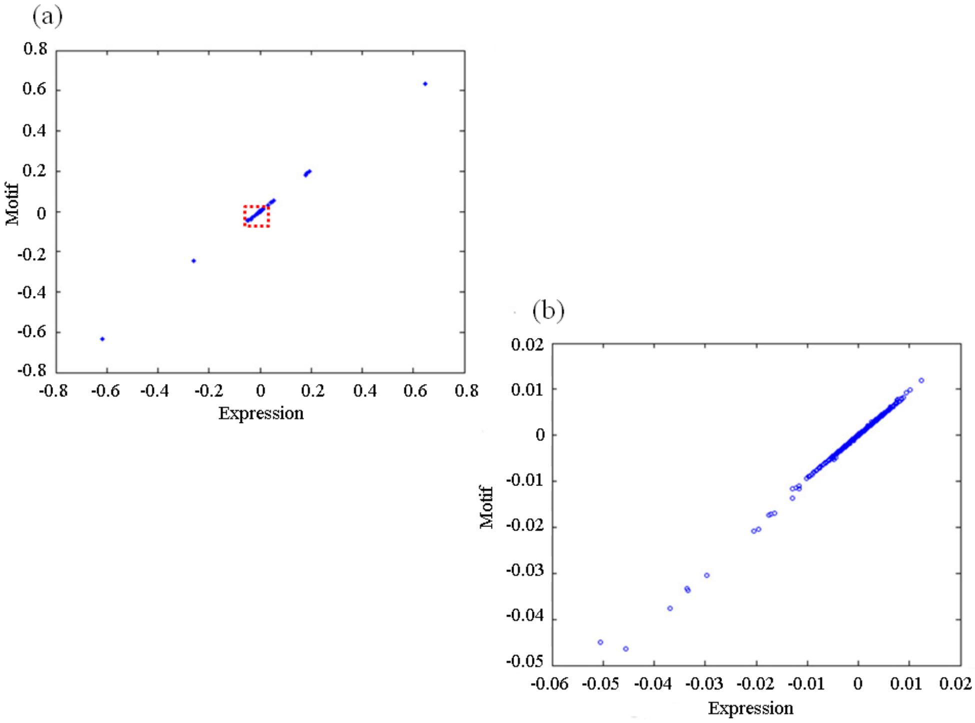


Figure 3.5: Correlation between expression profiles and motifs derived by using the raw upstream sequence data. The plot on (b) is an enlargement of the boxed area in (a).

raw sequences of 5-mers.

The 5-mer motif patterns with high weights are listed in Table 3.3. The 5-mer with the highest weight is 5'-GCGTG-3', which is similar to the MCB motif (5'-ACGCGT-3'). As described previously, MCB is an important motif involved in the cell cycle. The second-ranked sequence (5'-CGTGT-3') matched to the first five bases of the ALPHA2 motif sequence. From the second component, we also found several significant sequences, including a consensus sequence (5'-CGCGT-3') that

Table 3.3: High-scored motifs in the first and the second components using 5-mer raw upstream sequences.

| Sequence | Motif Description | Weight | Component | Rank |
|---|---|---|---|---|
| GCGTG | MCB (ACGCGT) | 0.079567 | 1 | 1 |
| CGTGT | MATalpha2 (CRTGTWWWW) | 0.075340 | 1 | 2 |
| CATGT | MATalpha2 (CRTGTWWWW) | 0.046299 | 1 | 12 |
| CCGGA | MCM1 (CCNNNWWRGG) | 0.044133 | 1 | 13 |
| TAAGG | MCM1 (CCNNNWWRGG) | 0.042387 | 1 | 15 |
| CCACG | SCB (CACGAAA) | 0.018992 | 2 | 4 |
| CGCGT | MCB (ACGCGT) | 0.017870 | 2 | 5 |
| GTGTT | MATalpha2 (CRTGTWWWW) | 0.016595 | 2 | 9 |

is identical to the MCB motif (5'-ACGCGT-3'). This further confirmed that the MCB motif affects gene expression in the cell cycle. Another interesting motif is 5'-CCACG-3', which is a sequence block with one base shift from the known SCB motif (5'-CACGAAA-3'). The SCB motif is a binding site of the SBF protein, which is a complex of SWI4 (a DNA-binding component) and SWI6 (a regulatory component) (Simon et al., 2001), and SBF is a major regulator in the G1/S transition.

### 3.3.3 Combinational effects of regulatory motifs

We searched the motif pairs that have synergistic or co-regulatory combination effects in the yeast cell cycle. The regulatory mechanisms of eukaryotes are highly complex since most genes are normally synergistically regulated by different transcription factors. Therefore, identifying the synergistic motif combinations can contribute to systematically understanding the regulatory circuit.

In the present study, using the kernel CCA we calculated the weight value for each motif pair of 42 known motifs. The heat map of weight values of all motif pairs is provided in Figure 3.6. Table 3.4 presents the top ten motif pairs with the highest weight values and with occurrence of more than ten in all the investigated upstream sequences. It also shows ECRScores which represent gene expression coherence. All these scores are relatively high compared to the previously identified synergistic motif pairs (ECRScores $> 0.075$). As shown in Table 3.4, the pair with the highest weight value is MCB-MCM1. According to a previous study, MCB and MCM1 were characterized as a significantly cooperative motif pair in the regulation of the cell cycle (Das et al., 2004). Other highly ranked pairs, such as ECB-ALPHA2 and MCM1-ALPHA2, are already known that they are required for transcriptional regulation of early cell cycle genes. MCM1 activates transcription of ECB (early cell cycle box)-dependent genes during M/G1 phase (MacKay et al., 2001), and the MCM1 protein can interact with the ALPHA2 factor regulating the expression of mating-type-specific genes (Keleher et al., 1989; Mead et al., 1996). These evidences support that two ALPHA2-related motif pairs act synergistically in the expressional regulation of the yeast cell cycle process. The REB1 motif, a binding site of REB1 protein, is frequently found among the pairs of motifs with the highest weights. The REB1 protein is an RNA polymerase I enhancer-binding protein and binds to genes transcribed by both RNA polymerase I and RNA polymerase II (Morrow et al., 1989). It is a general regulator rather than a condition specific one. Therefore, it is reasonable that this protein shows a high frequency in our results. REB1-SWI5, REB1-MCM1' and REB1-ALPHA1 motif pairs are already identified as acting synergistically in the yeast cell cycle regulation (Banerjee and Zhang, 2003; Tsai et al., 2005; Hvidsten et al., 2005). Most of our results are consistent with the previous reports. In addition, it's worth noting that several previously uncharacterized motif

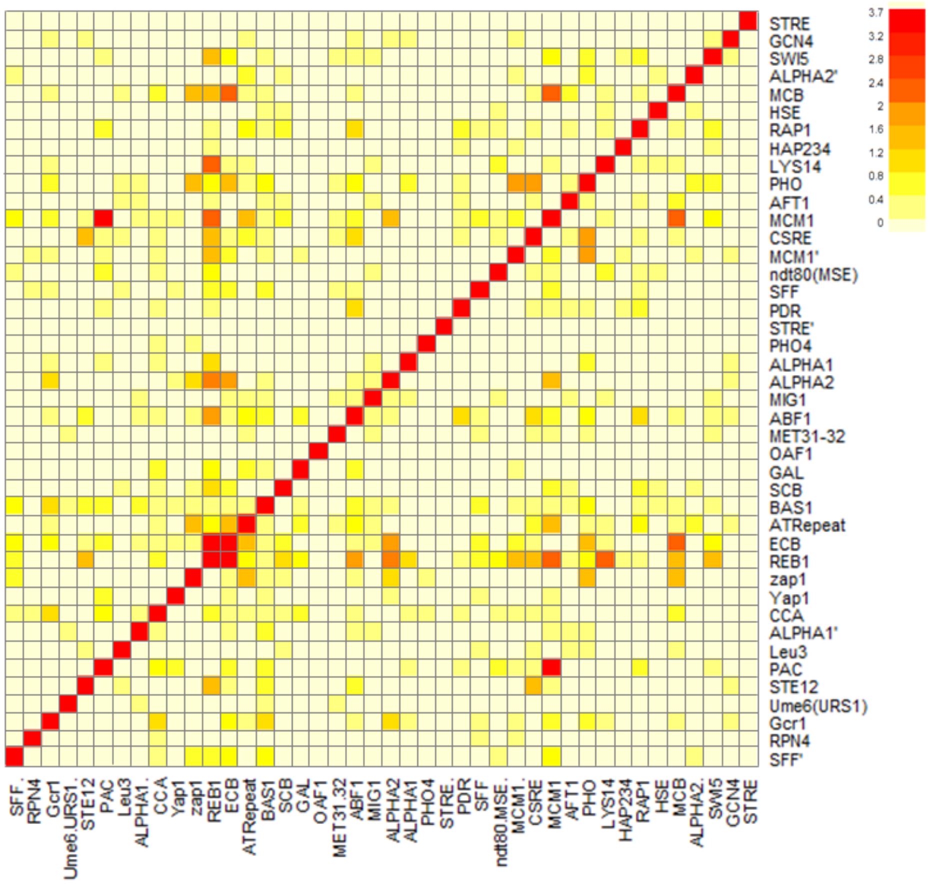


Figure 3.6: Heat map of weight values of motif pairs related to cell cycle regulation. Dark colors represent motif combinations of high weight values.

pairs were identified by our kernel CCA methods.

Table 3.4: The top 10 ranked motif pairs were extracted from the analysis of motif combination.

| Weight | Motif Pair | | ECRScore | Num. of ORFs |
|---|---|---|---|---|
| 2.5368 | MCB | MCM1 | 0.390 | 15 |
| 2.5018 | MCB | ECB | 0.439 | 12 |
| 2.0177 | PHO | MCM1’ | 0.088 | 17 |
| 1.848 | ECB | ALPHA2 | 0.088 | 14 |
| 1.7535 | MCM1 | ALPHA2 | 0.074 | 17 |
| 1.7263 | ATRepeat | MCM1 | 0.076 | 12 |
| 1.6995 | PHO | ECB | 0.127 | 11 |
| 1.6823 | REB1 | SWI5 | 0.099 | 14 |
| 1.6476 | REB1 | MCM1’ | 0.115 | 13 |
| 1.4256 | REB1 | ALPHA1 | 0.067 | 15 |

## 3.4 Discussion

We presented a novel method that can identify the candidate conditional specific regulatory motifs by employing kernel-based methods. The application of the kernel CCA enables us to detect correlations between heterogeneous datasets, consisting of upstream sequences and expression profiles. From a data-mining perspective, our work is regarded as a new approach for detecting important features from regulatory sequences and gene expression profiles. We demonstrated that major motifs in a specific biological process can be extracted by a CC score via modelling a close relationship between two datasets related to gene regulation.

As genome-wide datasets of various types become available, it’s important to analyze these datasets in an integrated manner (Kasturi and Acharya, 2005). It is possible to come up with novel biological hypotheses by integrating diverse biological resources generated for specific research purposes. In these aspects, the kernel CCA is regarded as a useful method that can extract the biological factors with significant roles by integrating different types of biological data. Many studies for identifying motifs have been based on sequence conservation or sequence characteristics, regardless of the biological processes. Therefore our method can be regarded as complementary approach in the analysis of gene regulation.

Our method found important motifs related to the cell cycle by using raw upstream sequences as well as known motif sets. In the present study we used the raw sequences of window size, $l$=5. If we enlarged the window size, the dimension for sequence features increased exponentially, whereas the frequency of motifs decreased. Although the window size used in our experiments was shorter than the length of several known transcription factor binding sequences, it was long enough to obtain worthwhile results.

In the future research, we will apply the proposed method to diverse gene expression datasets, especially cancer-related datasets. The cancer-related regulatory program can be elucidated by analyzing regulatory motifs from a set of enriched genes in the cancer transcriptome (Rhodes et al., 2005). Using the kernel CCA, a correlation analysis between regulatory sequences and the cancer transcriptome may directly catch regulatory motifs related to the abnormal gene regulatory program.

# Chapter 4

# Investigation of combinatorial effects of DNA methylation

## 4.1 Background

The addition of a methyl group to cytosine residues in the context of CpG dinucleotides (i.e., 5-methylcytosine) by the DNA methyltransferease (DNMT) enzymes is the most well studied epigenetic event. DNA methylation is known to play significant roles in many cellular processes, including embryonic development, genomic imprinting, X-chromosome inactivation, and preservation of chromosome stability. In addition, aberrant DNA methylation has been shown to disrupt many cellular processes and is frequently observed in most human diseases, including cancer (Suzuki and Bird, 2008; Robertson, 2005; Esteller, 2008; Keshet et al., 2006).

Methylation in CpG islands (CGIs), particularly in the promoter and first exon regions, is known to block genomic binding sites of activating transcription factors or other proteins and it is strongly associated with gene repression (Suzuki and Bird, 2008; Jones and Takai, 2001). In particular, the effect of DNA methylation on tumor suppressor genes (TSGs) has been extensively studied (Ueki et al., 2002). Transcriptional silencing of this key class of genes could contribute to defective regulatory processes in cancer, and the promoter CGI hypermethylation of TSG has been observed in a various types of cancers (Sakai et al., 1991; Merlo et al., 1995). However, few studies have examined the complex relationship between DNA methylation and gene expression on a genome-wide scale using accurate, high-resolution DNA methylation data.

Profiling of methylated CpG sequences is now possible by using next generation sequencing technologies and a number of recent studies have used high-throughput approaches to study DNA methylation (Chavez et al., 2010; Kim et al., 2011). Although generating enormous amounts (terabytes) of data is possible, single-base pair resolution of bisulfite-converted DNA is still costly and highly labor intensive. Recently, cost effective, genome-wide methylation approaches that do not rely on bisulfite-treated DNA have been developed, including methylation-sensitive restriction enzymes approaches (Zuo et al., 2009). One approach, the methylated-CpG island recovery assay (MIRA) (Rauch and Pfeifer, 2010) followed by sequencing (mCpG-seq), utilizes methylated-CpG-binding protein complexes with high affinity to methylated CpG dinucleotides in genomic DNA. Now a technique known as MBDCap-seq (Brinkman et al., 2010) is able to utilize double-stranded DNA, does not depend on the application of methylation-sensitive restriction enzymes, and generates DNA sequence variation data (Robinson et al., 2010).

The availability of high resolution DNA methylation and gene expression data on a genome scale now allows scientists to investigate the functional consequence of DNA methylation in various genomic regions, including CGIs which have been extensively investigated in the literature (Esteller, 2007; Bell et al., 2011; Pai et al., 2011). CGIs are often found near the promoter regions of genes and the CGI hypermethylation is known to have significant inhibitory effect on gene expression. In normal cells, CGIs are protected from methylation. However, hypermethylation of promoter CGIs of important genes, i.e. TSGs, is frequently observed in cancer cells (Sproul et al., 2011). In addition to CGIs, recent studies have reported that DNA methylation of other genomic regions can alter downstream gene expression. It was recently reported that methylation of CGIs near transcription start sites (TSSs) of genes (Sproul et al., 2011) or in CGI shores (Irizarry et al., 2009), regions about 2kb outside of CGIs, were both strongly associated with gene expression. In addition, a strong correlation between methylation in the first exon and expression of the corresponding genes was demonstrated (Brenet et al., 2011). Although these recent studies have clearly shown an association between DNA methylation at various genomic regions and gene expression, several questions remain to be answered. Specifically, in our study on the breast cancer cells, research questions are: How does DNA methylation in the different genomic regions contribute to gene expression? Are there subtype specific DNA methylation-gene expression patterns in breast cancer? Does the methylation of transcription factor binding sites impact transcription factor binding and subsequent gene expression?

To answer these questions, we used genome-wide profiling data from 30 breast cancer cell lines from the Integrated Cancer Biology Program (ICBP, http://icbp.nci.nih.gov/). We integrated MBDCap-seq methylation data and Affymetrix microarray gene expression data (Neve et al., 2006). The important goals of our study were:

1. Genomic studies have established major breast cancer intrinsic subtypes that show significant differences in incidence, survival and response to therapy (Koboldt et al., 2012). Basal-like breast tumors display aggressive clinical behavior and belong to the high-risk breast cancers that typically carry the poorest prognoses (Fadare and Tavassoli, 2008; Toft and Cryns, 2011). To

investigate whether phenotype specific methylation and expression patterns exist in the basal A, basal B, and luminal breast cancer molecular subtypes, we used an information-theoretic approach to identify genes with differentially methylated DNA regions and differential expression levels.

2. To perform an integrated analysis of DNA methylation and gene expression data on a genome-wide scale and to detect subtype-specific effects of DNA methylation in breast cancer cells. We examined relationships between DNA methylation and gene expression using step-wise analysis starting from genes whose expression was significantly altered in a particular subtype.

3. We used Pearson's correlation analysis and decision tree learning to investigate the effect of DNA methylation in various regions (CGIs, CGI shores, promoter regions, 1st exons, 1st introns, and 2nd exons) on the breast cancer subtype differential gene expression.

4. To investigate relationship between transcription factors and DNA methylation in promoter regions, we examined the relationship between DNA methylation specifically at transcription factor binding sites (TFBSs) and gene expression in the breast cancer molecular subtypes.

## 4.2 Materials and methods

### 4.2.1 Data

We prepared methylation and gene expression data from 30 breast cancer cell lines representing three tumor phenotypes found in patients (Neve et al., 2006): basal A, basal B, and luminal subtypes. Among 30 cell lines, 17 were basal-like and 13 were luminal-like subtypes (Table 4.1). The basal-like 17 cell lines were further subdivided into 7 basal A and 10 basal B subtypes.

Gene expression data from Affymetrix microarray experiments (Neve et al., 2006) was downloaded. Genome-wide methylation profiles were measured using the MBDcap-seq technique. The double stranded methylated fragments were sequenced and reads were mapped to the human reference genome. Methylation levels were calculated by measuring the density of the read coverage (Rao et al., 2013), as we have described previously.

The microarray gene expression data were processed and analyzed using R and Bioconductor. The expression values were centered by mean-adjusting each log abundance value (subtracting each value from the mean expression value in the cell line).

### 4.2.2 Profiling of DNA methylation patterns

To investigate DNA methylation characteristics across the 30 breast cancer cell genomes, methylation profiles were measured on ± 10 kb genomic regions around the TSS of each gene. We divided the genomic regions into bins with a size of 100 bases. DNA methylation levels were then measured as the number of mapped reads within each bin.

### 4.2.3 Identifying differentially methylated/expressed genes by information theoretic analysis

We identified differentially methylated and expressed genes in the three breast cancer subtypes using normalized entropy. Entropy is a measure of uncertainty, defined as follows:

$$\mathrm{H} = -\sum_{i=1}^{n} p_i \log p_i$$

Table 4.1: 30 Breast cancer cell lines and molecular subtypes

| Cell line | Subtype | Cell line | Subtype |
|---|---|---|---|
| BT549 | BaB | HCC1569 | BaA |
| HCC1937 | BaA | HCC1143 | BaA |
| HCC1428 | Lu | HCC202 | Lu |
| MDAMB436 | BaB | SUM185PE | Lu |
| 600MPE | Lu | HCC1500 | BaB |
| MDAMB231 | BaB | SUM225CWN | BaA |
| SKBR3 | Lu | MDAMB453 | Lu |
| SUM1315MO2 | BaB | SUM52PE | Lu |
| HSS78T | BaB | MCF12A | BaB |
| MDAMB157VII | Lu | HCC70 | BaA |
| HCC1954 | BaA | SUM149PT | BaB |
| GCC2185 | Lu | LY2 | Lu |
| MCF7 | Lu | BT20 | BaA |
| MCF10A | BaB | BT474 | Lu |
| SUM159PT | BaB | AU565 | Lu |

Lu: luminal; BaA: basal A; BaB: basal B

where $p_i$ denotes the probability of the state $i$, and $n$ is the total number of the states. In this study, the state $i$ is a cancer phenotype, i.e. $i = (basalA, basalB, Lu)$. For methylation profiles, the probability $p_i$ is measured by $t_{ji}/c_j$, where $c_j$ is sum of read counts for cell lines in a genomic region $j$ and $t_{ji}$ is sum of reads for a phenotype $i$ in the region $j$. For gene expression, $c_j$ is sum of expression values for cell lines in a gene $j$ and $t_{ji}$ is sum of expression for a phenotype $i$ in the gene $j$. The entropy $H$

achieves its maximum value when all states are equally probable, that is, it exhibits the lowest degree of uncertainty. If there is only one state, then the entropy $H$ is zero.

Normalized entropy is the ratio of entropy to maximum entropy as follows:

$$\mathrm{H}_0\left(x\right) = \mathrm{H}\left(x\right)/\mathrm{H}_{\max}$$

where $H_{max}$ is maximum entropy value where the probabilities are all equal.

We measured the normalized entropy and identified differentially methylated regions and differentially expressed genes. To avoid errors on the probability calculation, we introduced pseudo-probability to every zero-valued position.

### 4.2.4 Identifying downregulated genes in each subtype for integrative analysis

Genes differentially expressed in each different molecular subtype were further identified as follows. Suppose that $e_{gl}$ is an expression level of a gene $g$ in a cell line $l$. Since the cell line $l$ is clustered into a specific subtype $i$, we calculate the median values $Median(e_g, i)$ for the expression levels in each subtype $i$ per gene $g$. In this study, we measured three median value $Median(e_g, Lu)$, $Median(e_g, BasalA)$, $Median(e_g, BasalB)$ for each gene $g$.

If the median value $Median(e_g, i)$ of a gene $g$ in a type $i$ was significantly lower than those of other two types, we defined the gene $g$ as down-regulated in a specific type. In our study, log-ratio 1.5 was the criterion for significance.

### 4.2.5 Correlation between DNA methylation and gene expression

To investigate the relationship between methylation in various regions and gene expression in the 30 breast cancer cells, we examined methylation levels in gene

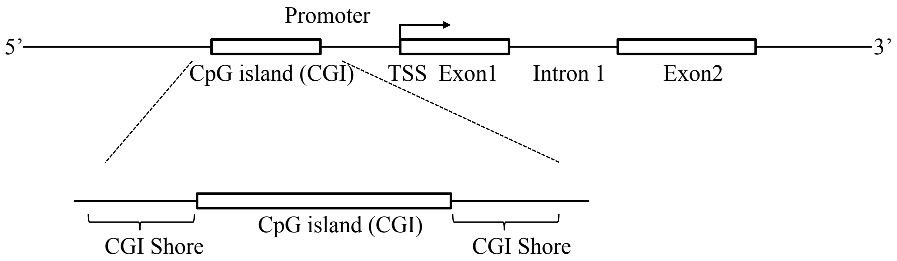


Figure 4.1: Genomic regions for studying DNA methylation profiles. A gene body is composed of promoter and coding regions including exons and introns. CGIs as well as these regions were studied for the effect of DNA methylation on gene regulation.

promoter regions (2kb upstream regions from TSSs), CGIs, CGI shores, the first and second exon and the first intron (Figure 4.1). The association between gene expression and methylation values of these datasets was measured by a Pearson's correlation coefficient. It was calculated on the paired data of a gene expression level and the methylation level in the genomic region.

### 4.2.6 Combinatorial effects of DNA methylation in various genomic regions

To identify which regions have dominant effects on downstream gene expression and also to investigate on the combinatorial roles of DNA methylation of the various genomic regions in each subtype, a decision tree was constructed using the methylation profiles in each region. For the learning purpose, a gene was an instance of data and gene expression was considered as a class variable, *i. e., up or down regulated genes*. The methylation value in each genomic region was an attribute. For binary classification, in training dataset of each subtype, the class values were discretized to high and low, *i. e., upregulated or downregulated genes.* If a gene was significantly downregulated in a subtype but the gene was upregulated in the other subtypes, the class values of the genes in the cell lines within the subtype were designated as low. For example, assume that the expression of a gene is significantly downregulated in Lu subtype. Then among 30 cell lines, 13 instances with Lu subtype are marked as low and 17 with the other types are high. The trees were built using REPTree in WEKA software (Hall et al., 2009).

### 4.2.7 Analysis of transcription factor binding regions possibly blocked by DNA methylation

For the integrative analysis of TFs, DNA methylation and gene expression, we used four datasets: gene expression, methylation profiles, cell specific DNA sequences and information for TF binding sites (TFBSs; TRANSFAC database (Matys et al., 2006)). We considered only downregulated genes in each subtype, as we were most interested in DNA methylation of TFBSs, possible interference on TF binding, and subsequent negative effect on gene expression. We referred to these downregulated as *target genes.* Differentially methylated genomic regions of the target genes were identified by statistical testing (t-test) of methylation levels at each 100bp-bin for the promotor regions. Cell-specific consensus sequences were computed by assembling short reads in the promotor regions of these genes. TFBSs were searched on the cell-specific consensus sequences corresponding to the hypermethylated bins, using ′minimize false positive′ option of the match tool in the TRANSFAC package (Kel et al., 2003).

Among the collected TFs that could be potentially blocked by TFBS methylation in the promotor region, we selected TFs whose expression levels were not significantly different in each phenotype (by t-test), as to exclude cases where the down-regulation of the target genes is as a result by difference in the expression levels of TF, an activator gene. In this way, we compiled cases where down-regulation of the target genes was due only to the hypermethylation in the promotor region, not other factors, such as the genomic sequences on the TFBSs and the expression levels of the TF.

## 4.3 Results

### 4.3.1 DNA methylation in 30 ICBP cell lines

We measured and compared the methylation density of 2kb promoter regions for all genes in 30 breast cancer cell lines. Figure 4.2 shows subtype-specific density plots of promoter regions, excluding unmethylated genes. Overall, the methylation density was similar in each subtype. We observe that the number of highly methylated ($>$ 50) promoter regions tended to be lower in BaB. The density of the regions whose methylation levels were over 50 was around 10% in Lu and BaA, but 4% in BaB.

Next, we investigated CGI methylation around each gene. CGIs are defined as regions of DNA of greater than 500 bp with a G+C equal to or greater than 55% and observed CpG/expected CpG of 0.65 (Takai and Jones, 2002). Using the position information of the CGIs from UCSC genome browser, we checked the methylation profile in the CGI near each gene. In the 30 breast cancer cell lines, approximately 47% of CGIs were methylated; however, distinct methylation density for each subtype was not apparent (Figure 4.3).

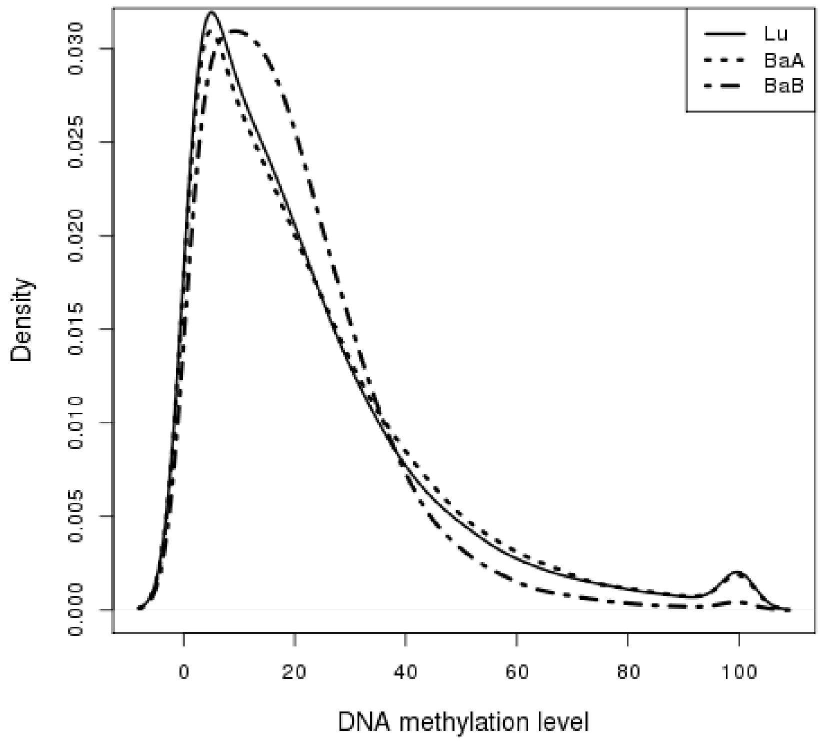


Figure 4.2: Methylation density of promoter regions in 30 breast cancer cell lines. Density was measured for each subtype. The methylation levels are on the x-axis and the y-axis is probabilistic density. Unusual bulbs around 100 on the x-axis were because methylation levels over 100 were truncated to 100. Lu, luminal; BaA, basal A; BaB, basal B.

### 4.3.2 Information theoretic analysis of phenotype-differentially methylated and expressed genes

To identify differentially methylated and expressed genes across the breast cancer genome, we measured normalized entropy. Lower entropy corresponded to genes

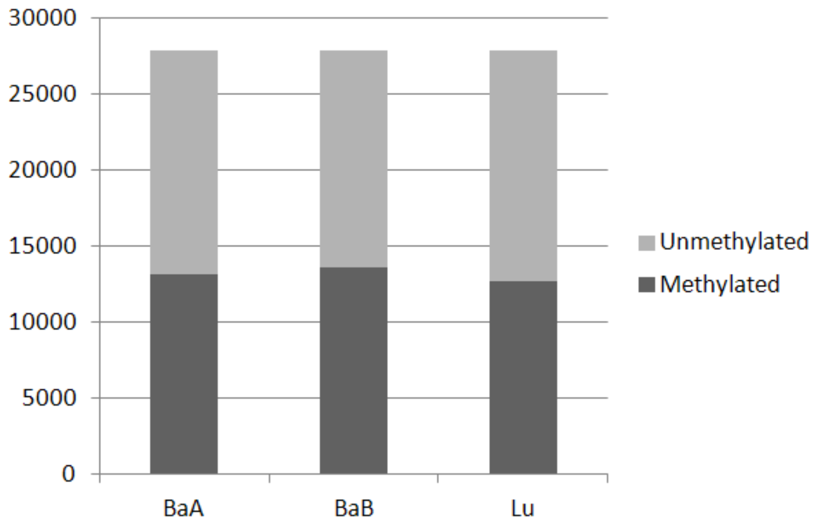


Figure 4.3: Average number of methylated and unmethylated CGIs in each cell. The unmethylated means that the mapped read count is zero in the CGI. BaA: basal A, BaB: basal B, Lu: luminal.

more differentially methylated or expressed in each subtype. First, we determined which genes were differentially methylated. Considering only genes with >3 mapped reads, there were 241 differentially methylated genes with the entropy threshold 0.2 and 564 differentially expressed genes with entropy threshold 0.5. Among these, only three genes were common to both the differentially methylated and expressed gene sets (Table 4.2) Thus, we concluded that separate analysis of differentially methylated and expressed gene sets based on information theory is not effective for the integrated analysis of methylation and gene expression, although these methods were effective to highlight genes and genomic regions that were different according to phenotypes.

Table 4.2: Genes that were both differentially methylated and expressed

| Gene Name | Description |
|---|---|
| PLA2G12A | phospholipase A2, group XIIA |
| FAT1 | FAT tumor suppressor homolog 1 |
| PARP8 | poly (ADP-ribose) polymerase family, member 8 |

### 4.3.3 Integrated analysis of DNA methylation and gene expression

To perform the integrated analysis of DNA methylation and gene expression, we used a two-step analysis process: (1) identify differentially expressed genes in each subtype, and (2) for each genomic region, test if there is a strong negative correlation between methylation level at the genomic region and the expression level of the gene.

To select differentially expressed genes in each subtype, we measured median values of expression levels for each of the three breast cancer phenotypes. If the median value of a gene in one subtype was significantly higher or lower than the median value in the other two subtypes, the gene was considered to be differentially expressed in a specific type. For such differentially expressed genes, variations of methylation levels were then investigated.

As DNA methylation is known to inhibit gene expression and an inverse correlation between the DNA methylation and gene expression has been shown to exist, we were most interested in a negative correlation between DNA methylation and gene expression for the integrated analysis. As an example, Caveolin 1, *CAV1*, represents a negative relationship between DNA methylation and gene expression (Figure 4.4). The *CAV1* gene has been shown by us and others to regulate breast tumor growth and metastasis and is overexpressed in basal-like subtypes (Sloan et al., 2004; Savage et al., 2007; Rao et al., 2013). *CAV1* expression levels were clearly different in each

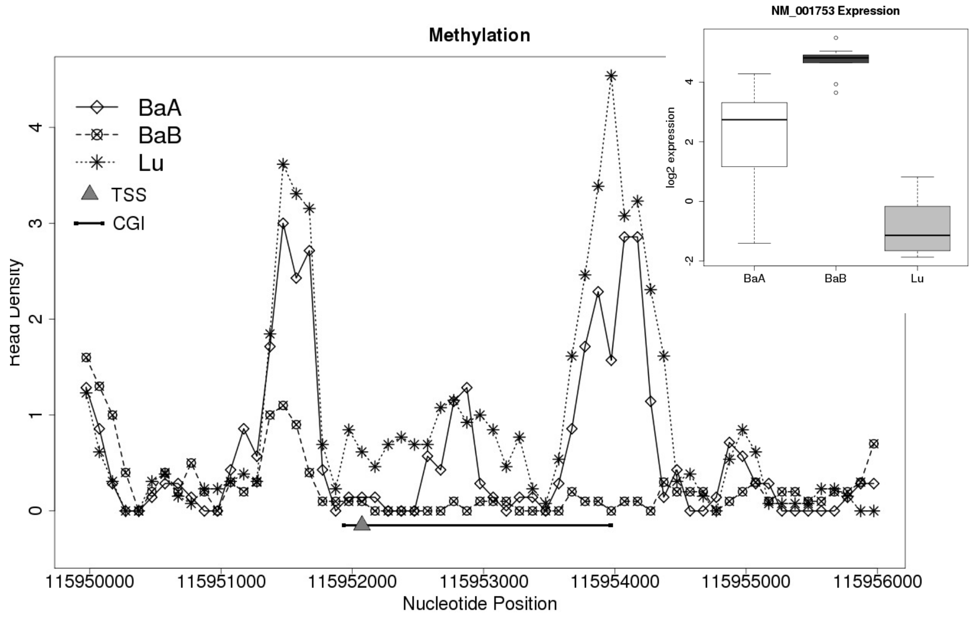
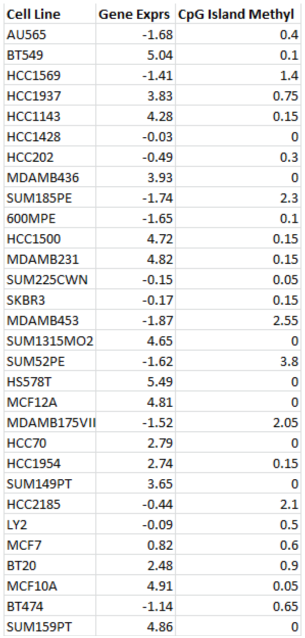


Figure 4.4: CGI methylation and gene expression of the CAV1 gene. Methylation and gene expression values from the 30 cell lines are grouped into luminal (Lu), basal A (BaA) and basal B (BaB) subtypes. **(a)** A plot showing the density of methylation in the CGI and shore regions located near the TSS of the CAV1 gene. The black bar shows the location of the CGI and the small orange triangle is the TSS. **(b)** A boxplot showing the expression of the CAV1 gene.

breast cancer subtype, higher in BaB subtypes and lower in Lu subtypes. However, when the DNA methylation profiles of the *CAV1* TSS and CGI were examined, methylation levels were significantly higher in the Lu compared to BaA and BaB. Furthermore, differential methylation of CGI shores, but not CGIs, significantly regulated *CAV1* expression, and breast cancer aggressiveness was associated with *CAV1*

(a)



| Cell Line | Gene Exprs | CpG Island Methyl |
|---|---|---|
| AU565 | -1.68 | 0.4 |
| BT549 | 5.04 | 0.1 |
| HCC1569 | -1.41 | 1.4 |
| HCC1937 | 3.83 | 0.75 |
| HCC1143 | 4.28 | 0.15 |
| HCC1428 | -0.03 | 0 |
| HCC202 | -0.49 | 0.3 |
| MDAMB436 | 3.93 | 0 |
| SUM185PE | -1.74 | 2.3 |
| 600MPE | -1.65 | 0.1 |
| HCC1500 | 4.72 | 0.15 |
| MDAMB231 | 4.82 | 0.15 |
| SUM225CWN | -0.15 | 0.05 |
| SKBR3 | -0.17 | 0.15 |
| MDAMB453 | -1.87 | 2.55 |
| SUM1315MO2 | 4.65 | 0 |
| SUM52PE | -1.62 | 3.8 |
| HS578T | 5.49 | 0 |
| MCF12A | 4.81 | 0 |
| MDAMB175VII | -1.52 | 2.05 |
| HCC70 | 2.79 | 0 |
| HCC1954 | 2.74 | 0.15 |
| SUM149PT | 3.65 | 0 |
| HCC2185 | -0.44 | 2.1 |
| LY2 | -0.09 | 0.5 |
| MCF7 | 0.82 | 0.6 |
| BT20 | 2.48 | 0.9 |
| MCF10A | 4.91 | 0.05 |
| BT474 | -1.14 | 0.65 |
| SUM159PT | 4.86 | 0 |

(b)

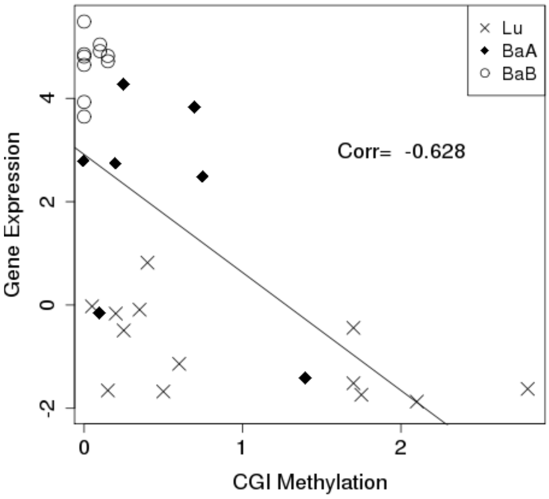


Figure 4.5: An example of the paired input data used to measure the Pearson correlation between gene expression and methylation. This paired data is for CAV1 gene. **(a)** Gene expression and CGI methylation across 30 cell lines. **(b)** Plot of gene expression profiles (y-axis) v.s. methylation levels (x-axis). Each pair in the cells is represented as a cross sign (Lu), a diamond (BaA) and a circle (BaB). A regression line is shown.

CGI shore methylation levels (Rao et al., 2013). The above negative correlation was measured by computing Pearson correlation coefficients. The Pearson correlation is measured by paired input data between DNA methylation profiles and gene expression levels across the 30 breast cancer cell lines. As an example, a correlation coefficient from CGI methylation and gene expression levels was calculated across

30 cell lines (Figure 4.5). The scatter plot for *CAV1* gene shows that gene expression and CGI methylation levels were negatively correlated.

We measured the methylation correlation for various genomic regions of downregulated genes in Lu and BaB subtype (Figures 4.6 - 4.7). Since only two genes were detected as downregulated in BaA subtype, the correlation results for BaA subtype were not included. Interestingly, when methylation in promoter regions was considered, several genes showed a clear negative correlation at the proximal regions of TSSs. Figure 4.6 is heatmaps that visualize promoter region methylation and downstream gene expression (light red colors mean that two vectors (methylation profiles and expression levels) were highly negatively correlated and bright green were positively correlated). In both Lu and BaB subtypes, strong negative correlations were observed in promotor regions, and methylation in the promotor regions near TSS showed strongest negative correlations. However, there were significant differences in promotor methylation patterns in Lu and BaB subtypes. In Lu subtypes, weaker negative correlations were observed at genomic regions further away from TSS. On the contrary, in BaB subtypes, consistently strong negative correlations were observed in entire promotor regions. This result implies that the DNA methylation on the promoter region has stronger epigenetic inactivation in Basal-like subtypes and the methylation of this regions may contribute to breast cancer progression.

Moreover, in most genes, first exon and CGI methylation levels were negatively correlated with expression levels (Figure 4.7). From the multi-exon genes, we measured correlation coefficients between the DNA methylation profiles for each exon and intron, and the expression level of the corresponding gene. A clear negative correlation was observed in the first exon, but this was not the case for second exons and first introns, a result consistent with a previous study showing that first exon methylation was closely associated with low gene expression (Brenet et al., 2011).

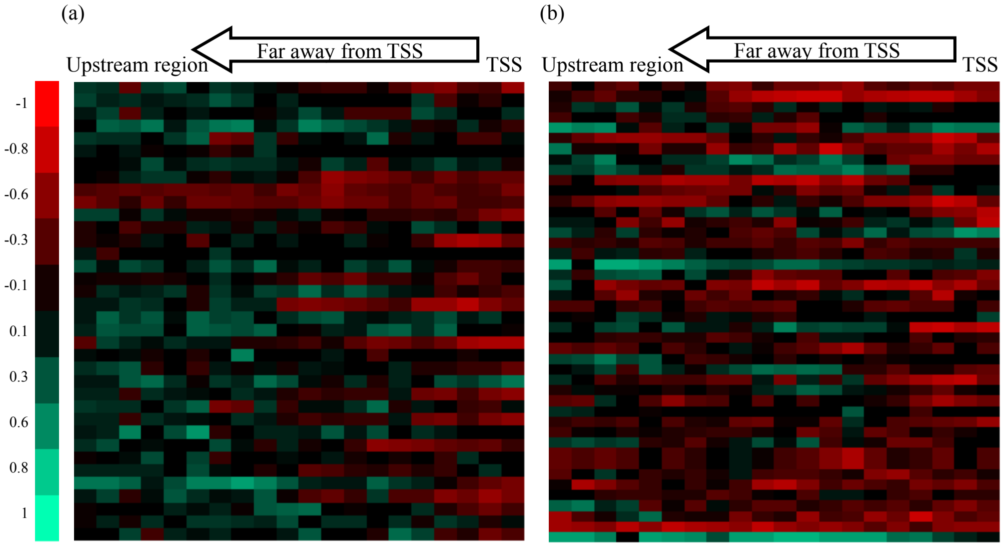


Figure 4.6: Correlation between promoter region methylation profiles and expression levels of genes downregulated in (a) Lu and (b) BaB subtypes. Unmethylated genes in the whole promoter region of 30 cell lines were excluded. Light red color was used for negative correlation and light green for positive correlation. Columns from right to left denote positions getting away from TSS. Each row is a downregulated gene in the subtype.

When we examined CGIs and CGI shore regions, negative patterns were also apparent. CGI and CGI shore DNA methylation levels were negatively correlated with gene expression levels in most genes, but in CGIs, much stronger relationships were shown in our datasets.

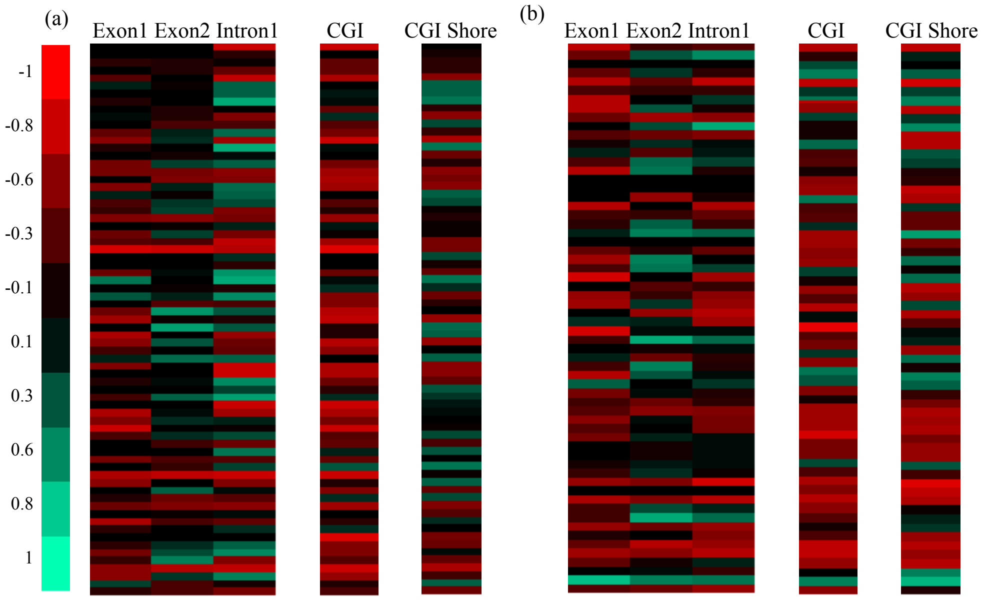


Figure 4.7: Correlation between methylation profiles on CGI, CGI shore, intron, and exon regions and expression levels of genes down-regulated in (a) Lu subtypes and (b) BaB subtypes.

### 4.3.4 Investigation of the combinatorial effects of DNA methylation in various regions on downstream gene expression levels

As DNA methylation occurs in many genomic regions, it was of interest to examine the effect of the various regions on downstream gene expression, particularly which regions may have a dominant effects on gene expression and whether the effects of the regions were similar in each subtype. Towards this goal, we performed a comprehensive study using six distinct genomic regions: promoter regions, CGIs, CGI shores, first and second exons, and first introns. Using the DNA methylation profiles in these regions, we performed a machine learning analysis.

The decision tree is a classification method that uses conjunctions of features for predicting target values in a tree-like hierarchical decision process. As decision tree learning identifies the most informative attributes for classification, this approach was used to discover regions with dominant and combinatorial effects on expression levels. We normalized the methylation levels of each region in a gene by adjusting the scale, then carried out the decision tree analysis.

The decision tree was constructed with a constraint of a maximum tree depth of three excluding leaf nodes, and in this case, the classification accuracy for genes, downregulated in Lu subtype, was 0.649 in a 10-fold cross validation (Figure 4.8 (a)). In the decision tree, the right-most branch means that the nodes in this branch were hypermethylated, and the left-most that the regions were hypomethylated. Consistent with the correlation analysis, CGIs were the most informative feature.

In the BaB subtype whose classification accuracy was 0.746 with the same maximum depth, the promoter regions and the first exons had combinatorial effects on gene expression (Figure 4.8 (b)). In the left branch of the decision tree where TSS1001-2000 were hypomethylated, it is intuitive that genes were unregulated. However, in the left branch, when TSS1-1000 was hypermethylated and also the first exons were hypermethylated, genes were down regulated. Note that TSS1001-2000 region had the dominant effect on the gene expression in the BaB subtype. This was consistent with our previous correlation analysis showing a clear negative correlation in much broader regions (Figure 4.6). Since CGI overlaps the first exon or promoter regions, we carried out the analysis again by separating into two cases: (1) CGI overlaps with the regions and (2) CGI does not overlap with the regions. Even when we separated CGI overlapping cases, the dominant factors (CGI for the Lu subtype and TSS1001-2000 for the BaB subtype) remained the same as when

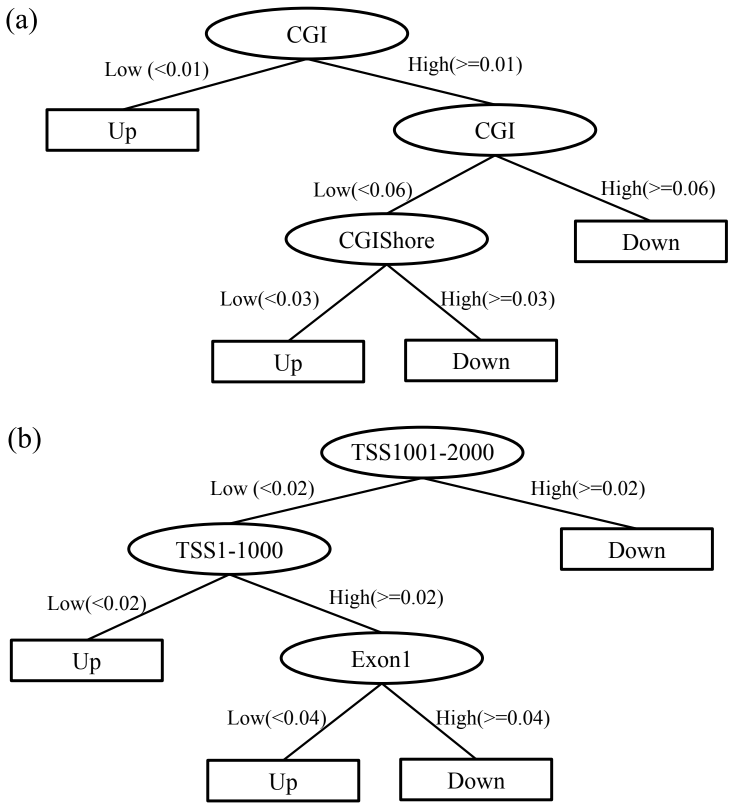


Figure 4.8: Decision tree analysis with downregulated genes in **(a)** Lu subtypes and **(b)** in BaB. The attributes are represented by circles, in where Exon1 is the first exon and CGIShore means 2kb outside region from CGI. TSS1-1000 means 1 to 1000 bp upstream region from TSS and TSS1001-2000 means 1001 to 2000 bp upstream. The Down in leaf nodes (rectangular boxes) means the gene is downregulated and Up means upregulated.

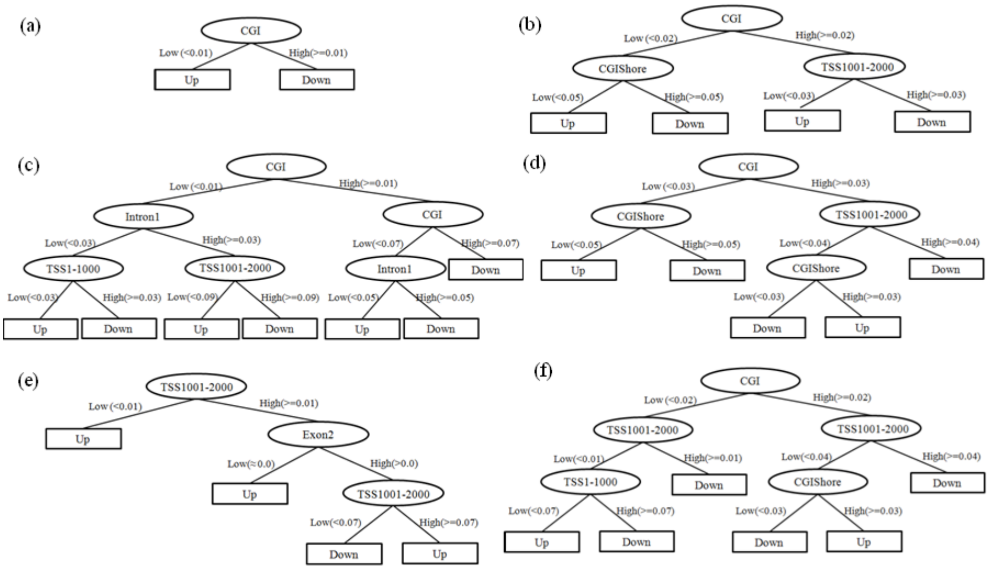


Figure 4.9: In case of down-regulation in Lu subtype, decision tree analysis separated by genomic regions of CGI. (a) Overlap with the first exon (The classification accuracy, Acc. is 0.737), (b) Nonoverlap with the first exon (Acc. is 0.590),(c) Overlap with TSS1-1000 (Acc. is 0.687), (d) Nonoverlap with TSS1-1000 (Acc. is 0.644), (e) Overlap with TSS1001-2000 (Acc. is 0.644) and (f) Nonoverlap with TSS1001-2000 (Acc. is 0.644).

we did not separate CGI overlapping cases. The decision trees when we did not separate CGI overlapping cases were presented in the main text (Figure 4.8) and the decision trees when we separated CGI overlapping cases were presented in Figures 4.9 and 4.10. The decision tree results suggest that altered gene expression in the two subtypes is associated with not only different promoter methylation profiles but also different combinatorial effects in various genomic regions.

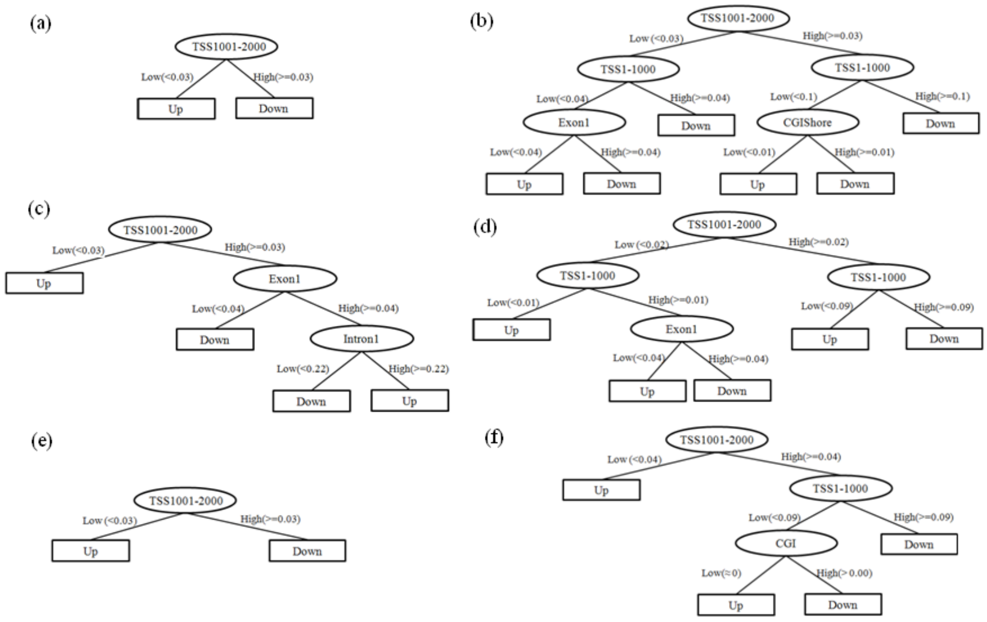


Figure 4.10: In case of down-regulation in BaB subtype, decision tree analysis separated by genomic regions of CGI. (a) Overlap with the first exon (Acc. is 0.773), (b) Nonoverlap with the first exon (Acc. is 0.760),(c) Overlap with TSS1-1000 (Acc. is 0.810), (d) Nonoverlap with TSS1-1000 (Acc. is 0.708), (e) Overlap with TSS1001-2000 (Acc. is 0.824) and (f) Nonoverlap with TSS1001-2000 (Acc. is 0.741).

### 4.3.5 Integrative analysis of transcription factors, DNA methylation and gene expression

We next sought to investigate the effect of DNA methylation on the interaction between TF and DNA, *i.e.* binding of a TF to the promotor region of a gene. To investigate this important concept, we developed a rigorous data mining protocol to compile a list of TF that are potentially blocked by DNA methylation. The schematic

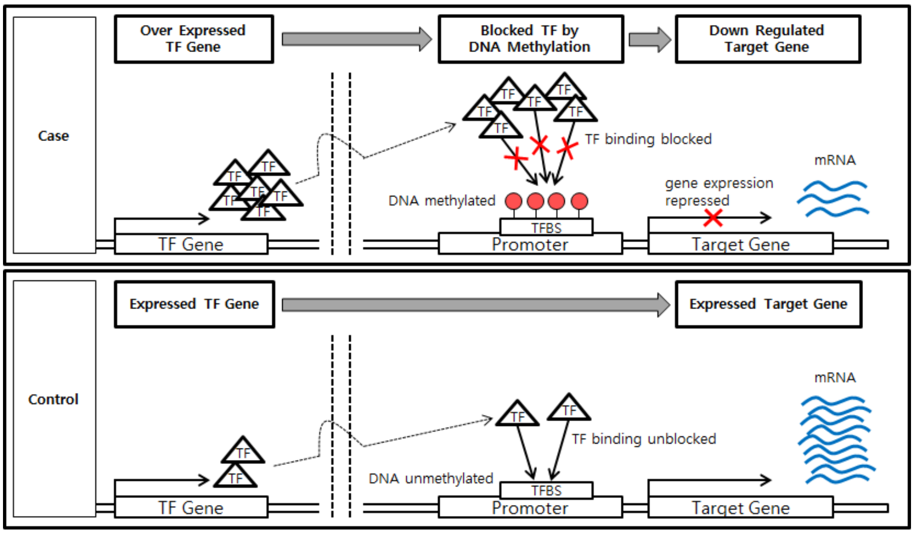


Figure 4.11: Schematic overview of the phenotype-comparative analysis for interference of TF binding by DNA methylation resulting in the suppression of downstream gene expression

overview of the protocol is illustrated in Figure 4.11.

We first identified differentially methylated genes among the downregulated genes, 60 genes in BaB subtype and 52 genes in Lu subtype. Based on the results of the one side standard t-test with a criterion for being significant as p-value<0.005, we observed eight genes with significant hypermethylation in at least one 100bp-bin as follow: *CDH1*, *CLDN4*, *ESRP1*, *GRHL2*, *KRT19*, *PRR15L*, *AKR1B1*, and *PLOD2*. Figure 4.12 shows the promotor regions of the eight genes that are differentially methylated according to the p-values.

Next, for the hypermethylated regions of the eight genes, we generated cell line-specific consensus sequences by assembling short reads mapped to the regions and

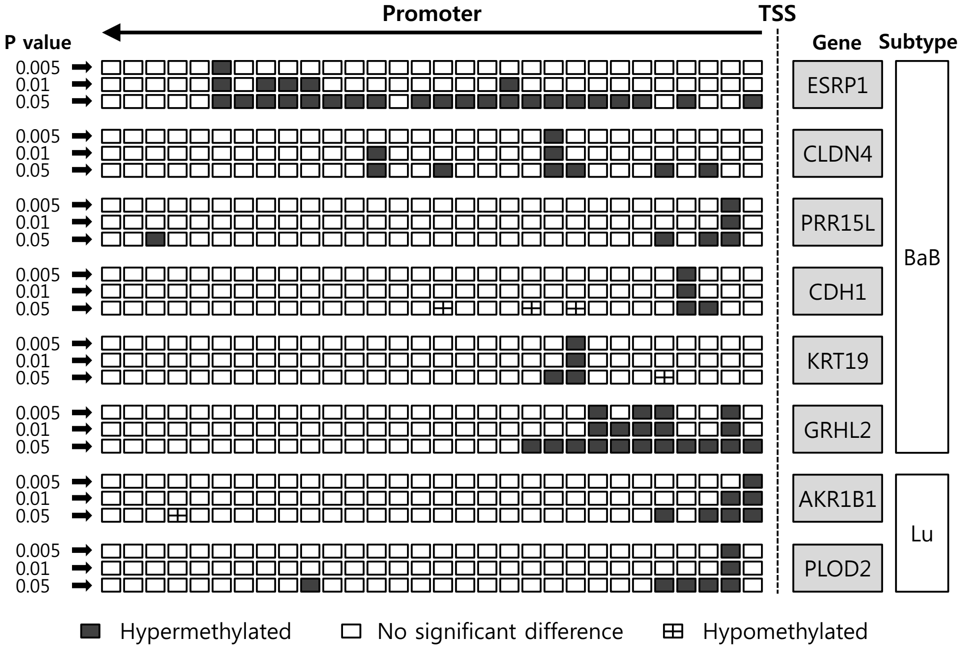


Figure 4.12: Differentially methylated promoter regions of down-regulated genes. Each rectangle in the upstream region means a 100bp-bin.

searched candidate TFs which can be bound to these consensus sequences by match tool (Kel et al., 2003) on the consensus sequences. To exclude the possibility that higher expression of an activator gene might result in upregulation of target genes, we discarded TFs whose expression levels were significantly different across cell lines of different phenotypes.

Table 4.3 summarizes the final selection of TFs and their target genes. TFs appeared in at least 50% of cell lines of the same phenotype (*TFBS Support Rate* in the table is percentage of the number of TF-containing cell lines). Interestingly the genes *CDH1*, *ESRP1* and *GRHL2* have been shown to play critical roles in epithelial-mesenchymal transition (EMT), a process associated with metastatic events in cancer and also highly relevant to tumor progression (Thiery, 2002; Thiery et al., 2009). Lombaerts *et. al.* (Lombaerts et al., 2006) reported that *CDH1* is downregulated by promoter methylation and related to EMT in breast cancer cell lines. A study by Dumont *et. al.* (Dumont et al., 2008) showed that the induction of EMT was accompanied by repression of *CDH1* expression and subsequent DNA hypermethylation at its promoter in basal-like breast cancer. Additionally, recent studies showed that *GRHL2* and *CDH1* in human breast cancer cells were highly correlated and suppressed EMT by repressing expression of the *ZEB1* gene (Xiang et al., 2012; Cieply et al., 2012). *ESRP1* was shown to regulate a switch in CD44 alternative splicing, an event required for EMT and breast cancer progression (Brown et al., 2011). Moreover, there might be potential interplay between target genes. Over-expression of *GRHL2* up-regulated *ESRP1* expression (Xiang et al., 2012), and *GRHL2* was shown to be essential for adequate expression of the *CDH1* and *CLDN4* (Werth et al., 2010). Thus, our approach may be useful to elucidate cell-specific regulatory mechanism using the genome-wide methylation data from the MBDCap-seq.

## 4.4 Discussion

Recent developments in sequencing technologies have made it possible to analyze genome-wide DNA methylation profiles at high resolution. Although altered DNA methylation patterns are a hallmark of cancer, and promoter CGI hypermethylation is known to repress gene expression, only a few studies have examined DNA methylation-gene expression relationships using genome-wide integrated analyses (Ruike et al., 2010; Fang et al., 2011; Sun et al., 2011). Several researchers have attempted to investigate the association of the DNA methylation with the molec-

Table 4.3: Downregulated target gene with transcription factor binding sites on hypermethylated region

| Target Gene | Binding TF | TFBS Support Rate |
|---|---|---|
| CDH1 | SMAD1 | 100.0 |
| CDH1 | FOXO1 | 100.0 |
| CLDN4 | CEBPA | 62.5 |
| CLDN4 | CEBPB | 62.5 |
| CLDN4 | CEBPD | 62.5 |
| CLDN4 | CEBPE | 62.5 |
| CLDN4 | CEBPG | 62.5 |
| ESRP1 | CUX1 | 90.0 |
| GRHL2 | PDX1 | 100.0 |
| KRT19 | PAX6 | 60.0 |
| PRR15L | IKZF1 | 50.0 |
| AKR1B1 | E2F1 | 91.7 |
| PLOD2 | PAX3 | 100.0 |

ular subtypes in breast cancer cells (Bloushtain-Qimron et al., 2008; Holm et al., 2010). However high resolution sequencing data were not used in those studies. To better understand the relationship between DNA methylation and gene expression in breast cancer molecular subtypes, we used next generation DNA methylation sequencing data and gene expression profiles for 30 ICBP cell lines representing molecular subtypes of the disease to perform a systematic analysis.

We first compared genome-wide methylation profiles of breast cancer phenotypes. Although overall DNA methylation profiles were similar in Lu, BaA and BaB, specific genomic regions were differentially methylated among the three subtypes. We then explored computational methods for integrating DNA methylation and gene expression data and started with differentially expressed genes for discovering genes whose expressions were influenced by DNA methylation.

DNA methylation of different genomic regions has recently been associated with altered expression of downstream genes. To better understand possible transcriptional regulatory roles of DNA methylation, we performed a comprehensive study considering distinct genomic regions: CGIs, CGI shores, promoter regions, 1st exons, 1st introns, and 2nd exons. Based on Pearson's correlation coefficients, we verified that the DNA methylation of several genomic regions including CGI and CGI shores were negatively correlated with downstream gene expression.

To investigate combinational effects of DNA methylation in these regions and to identify subtype-specific events, we applied a decision tree algorithm using genes downregulated in each subtype. Interestingly, we found potential combinatorial effects of the first exon methylation and promoter region methylation on the downstream gene expression (BaB subtype) and potential combinatorial effects of CGI methylation and CGI shore methylation (Lu subtype). As gene expression is regulated by many factors, it is difficult to predict gene expression levels using only the DNA methylation profiles. However, the classification accuracy was significantly high enough to elucidate the contribution of each genomic region and combinatorial effects of the regions. We showed that DNA methylation had combinatorial roles on gene expression and the effects of DNA methylation in each genomic region differed among the subtypes. Moreover, our studies further imply that the aberrant DNA methylation state of the TF-associated regions could be another contributing factor to gene repression, a subject of future experimental validation.

It is now well established that different gene expression patterns contribute to breast cancer heterogeneity (Koboldt et al., 2012). In the current study, our integrated analysis further demonstrates that methylation status of different genomic regions may play a key role in establishing transcriptional patterns in three molecular subtypes of human breast cancer. Understanding the functional impact of distinct regions of DNA methylation on gene expression patterns may provide additional insight into breast cancer progression and response to therapy, both critical for improving management of the disease.

# Chapter 5

# Detecting multiple SNP interaction via evolutionary learning

## 5.1 Background

Genome-wide association study (GWAS) examines genetic variations on the whole genome of individuals and investigates how the variants frequently occur in population with a particular phenotype such as disease. The main purpose of the GWAS is to identify the genetic variations which influence to phenotypic changes or are susceptible to diseases. One of the most popular variants to use in the GWAS is single-nucleotide polymorphism (SNP). SNPs were relatively easy to be identified, and many people believed that the cause of disease would be discovered by the variants. In reality, there have been a lot of research to capture the genetic variants which are statistically associated disease or traits, and as a result of GWASs, it has been reported that hundreds of loci are associated for more than 70 common diseases and traits (Donnelly, 2008).

However, comprehensive understanding for the relationship of genotypes to phenotypes, is still challenging. The complex traits including cancers and diabetes are believed to be affected by the interactions of multiple genetic factors (Cordell, 2009). In many cases, the single genetic variants did not fully explain a cause of the complex disease.

To understand the complexity of mapping from genotype to phenotype, many researchers have focused on genetic interaction and relationships of more amount of variants, instead of a single genetic marker (Heidema et al., 2007; Cordell, 2009). Especially, machine learning approaches could be a useful solution of the problem (Szymczak et al., 2009; Moore et al., 2010). For example, logic regression and decision trees could be applied for the analysis of the interaction of variants (Ruczinski et al., 2004; Fiaschi et al., 2010). Another widely used technique is MDR (multifactor dimensionality reduction) approach (Ritchie et al., 2001) which has been developed with the idea of CPM (combinatorial partitioning method). However, these have limitations in efficiently handling higher order interactions from a large number of SNPs.

Here, we address the multiple SNP associations to disease, by the construction of a classifier based on evolutionary learning. One of the important steps to improve the performance of a classifier is to identify the informative feature sets. Especially, in the association study, the number of features is very high, and in the case of concerning all of the multiple combinations of the attributes, most of computational learning algorithms might fail to efficiently control the large-scale datasets. We introduce a concept of evolutionary learning to identify higher-order combinatorial features which are relevant to the class discrimination, from the combinatorial search space. Generally, evolutionary learning well-approximates solution to complex problems which are difficult to optimize mathematically. For the genetic association studies, several research has been accomplished by the evolutionary learning, and showed that it could be applied successfully (Namkung et al., 2007; Moore and White, 2007; Nunkesser et al., 2007; Clark et al., 2008; Yang et al., 2010).

We propose a method to find association of multiple SNPs and a disease, and to predict a disease by the variant information. Firstly, we applied the approach to a simulation data and verified the approach could be useful to find the SNP interactions. After that, we identified the combinatorial effects of multiple SNPs on T2D in Korean population. In our evolutionary algorithm, a single individual is encoded by the form of explicit rules which are formulated for certain values of the attributes, and the whole population evolves to the final rule-set with a good fitness. In the learning process, the evolutionary computation can solve the problem efficiently by avoiding exploring the whole search space and leading to identify higher-order SNPs with strong association to a phenotype. The resulting rule set is able to correctly recognize instances and discriminate them to target concepts well. As a result, the model can classify the instances by combination of the survived rules after evolutionary learning, and the rules can be considered as informative multiple factor interactions associated to a disease.

## 5.2 Materials and methods

### 5.2.1 Identifying higher-order interaction of SNPs

The evolutionary computation approach, particularly learning classifier system (LCS) has successfully applied to induce a set of classification rules in a given environment (Bernado-Mansilla and Garrell, 2003; Sigaud and Wilson, 2007; Fernandez et al., 2010). The LCS searches the space of possible rules, guiding the search for better rules by evolutionary computing techniques. Our main goal is similar to the technique. We construct an evolutionary learning method guided by a gradient descendent algorithm, to induce a set of classification rules from SNP data with complex traits. The detail explanation follows.

#### Structure of the individuals

Suppose that $X = \{X_1, X_2, ..., X_n\}$ is a dataset of $n$ samples, and each sample $X_i$ is composed by $k$ features, that is, SNP loci, and class value $y_i \in \{normal, disease\}$. The input value of each feature in the SNP data can take one of the following three states: (1) homozygous major form, (2) heterozygous, and (3) homozygous minor form. The structure of the individuals are expressed as a combination of SNP information. For example, an individual is represented from the conjunctive form of the multiple SNP association as follows:

$(SNP_1 = 3) \bigwedge (SNP_2 = 2) \bigwedge (SNP_3 = 2)$

It means that the $SNP_1$ is hetero, $SNP_2$ is homo minor form and $SNP_3$ is also homo minor form.

### 5.2.2 Algorithm Description

The algorithms steps are summarized in Table 5.1 and Figure 5.1. More detail is given on individual steps in following subsections.

#### Initialization

In the evolutionary learning, population is defined by a set representing higher-order interaction among SNPs. The initial population is consisted by individuals randomly generated with chromosome length $l$. The population size $s$ is decided empirically and the initial weight $w_j$ of the individual $j$ $(0 < j < s)$ is randomly assigned with

Table 5.1: Overall learning procedure

**Main Learning Procedure:**

1. Randomly generate a population and initialize $s$ individuals with weights $w$s. The length of chromosome $l$ is user-specified. The weight (fitness) $\mathbf{w}$ is randomly initialized with a small value.

2. Train the weight value of each individual iteratively using instances. The weight values are updated and assigned by a gradient-descent algorithm. The learning procedure in step 2 is terminated when the weights are converged after repetition of a number of epoch.

3. The evolutionary process begins. Remove individuals with worst fitness from population. The individual is worse as its fitness is closer to zero. Theses are replaced by newly generated individuals. The offsprings are reproduced by one of four ways in user-specified proportion.

   (a) Inherit $r$ individuals whose $e_j$ is -1. (elitism)

   (b) $\alpha$ individuals should be generated by the crossover operator. By selection strategy (ranking selection), select two individuals and crossover them.

   (c) Mutate $\beta$ individuals in the parents.

   (d) Randomly generate $s - r - \alpha - \beta$ individuals

4. Go to Step 2 until convergence after the number of generation.

$r$ is a parameter for the number of removing individuals

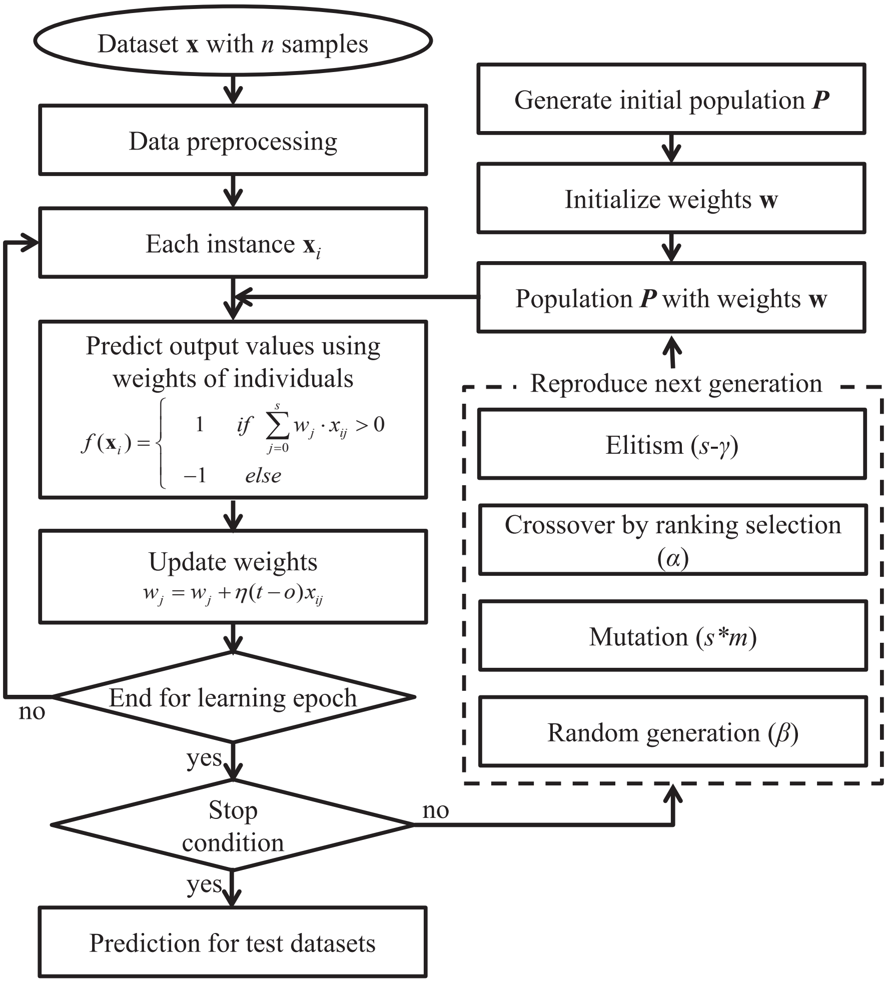

Dataset **x** with $n$ samples
Data preprocessing
Each instance $\mathbf{x}_i$
Predict output values using weights of individuals
$f(\mathbf{x}_i) = \begin{cases} 1 & if \sum_{j=0}^{s} w_j \cdot x_{ij} > 0 \\ -1 & else \end{cases}$
Update weights
$w_j = w_j + \eta(t - o)x_{ij}$
End for learning epoch
no
yes
Stop condition
no
yes
Prediction for test datasets
Generate initial population $\boldsymbol{P}$
Initialize weights **w**
Population $\boldsymbol{P}$ with weights **w**
Reproduce next generation
Elitism ($s$-$\gamma$)
Crossover by ranking selection ($\alpha$)
Mutation ($s*m$)
Random generation ($\beta$)

Figure 5.1: Flow chart for our evolutionary learning method. The most fitable individual is searched by the iterative learning.

a small value (-1 $< w_j <$ 1).

### Weight Update and Evaluation

Each individual has a weight value which means how informative the chromosome is to classify the samples. That is, the weights for individuals are considered as their fitness and the bigger weight on an individual mean mores informative to classify the instances. To determine and update the fitness for each individual, we introduce a gradient descendant rule as follows:

$$w_j = w_j + \eta(t_i - f(\mathbf{x}_i))m_{ij}, \tag{5.1}$$

where $w_j$ is a weight value for $j$-th individual and $t_i$ is a target class in the $i$-th training instance. $m_{ij}$ is a variable whether the all values of attributes within the $j$-the individual is matched to those in the $i$-th instance.

$$m_{ij} = \begin{cases} 1, & \text{if all values are identical} \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

$f(\mathbf{x}_i)$ is a predicted output value of the $i$-th training instance by our model and determined as follows:

$$f(\mathbf{x}_i) = \begin{cases} 1, & \text{if } \sum_{j=0}^{s} w_j \cdot m_{ij} > 0 \\ -1, & \text{else} \end{cases} \tag{5.3}$$

The difference between the predictions and the target values specified in the training sequence is used to represent the error of the current weight vector. The target function is optimized to minimize the classification error. The weight values are evaluated against a sequence of training samples and are updated to improve the classification accuracy. The weight update processes are repeated until it would be converged after the number of epoch.

Using the learning scheme, we find most informative individuals for classification, that is, the absolute value of their weights is large.

### Removing and Reproduction

During each successive generation, a proportion of the existing population is selected to be survived in the next generation. We adopted individual replacement strategy in each generation during the evolution processes. Basically, the highly weighted individuals should be selected and the others dismissed. It is a similar concept with elitism in a conventional genetic algorithm. We measure the fitness of each individual and preferentially select $r$ best solutions. The $r$, the number of individuals to be survived, is determined by a threshold $\theta$.

$$e_j = e(w_j) = \begin{cases} -1, & \text{if } |w_j| < \theta \\ 1, & \text{otherwise} \end{cases} \tag{5.4}$$

, where $|\cdot|$ means a absolute value. Then, individuals whose $e_j$s are 1, is survived and the $s$ - $r$ individuals are removed. After that we generate new individuals as much as removed in the step.

$s$ - $r$ individuals are reproduced by three ways in the next generation. The first is random generation. As similar to the the process for making initial population, we can construct new individuals randomly. Another ways are bring from conventional genetic operators, crossover and mutation. We select two individuals by ranking selection and can recombinate them in a random position. $\alpha$ and $\beta$, the number of individuals to be generated by crossover and mutation, respectively, are determined as follows:

$$\alpha = \lambda(s - r) \tag{5.5}$$

$$\beta = s - r - \lambda(s - r) \tag{5.6}$$

, where $\lambda$ is a crossover rate. For the mutation operator, there exists two kind of alteration. We choose $\kappa$ individuals and substitute a gene to another gene. There is the other base mutation rate $\rho$. It change a value of a variable in a selected individual for mutation to other one, so it help to be exploited in the search space by giving a variation in the combinatorial factors of the individual.

$$\beta = \mu + \kappa = \rho\beta + (1 - \rho)\beta \tag{5.7}$$

, where $\rho$ is the base mutation rate, and $\mu$, the number of individuals to be occurred base mutation, is determined by the $\rho$ value.

#### Iterative Learning Procedure

To select interesting rules from population, that is, the sets of the possible rules, we iteratively reproduce the individuals in progress on generation. The individuals are required to satisfy the specified fitness function and are survived only if they are adapted in the environments, that means they are how much informative to classify the training dataset.

By repeating the procedures until convergence (Table 5.1), the model can classify the normal and disease samples well, and identify higher-order interactions of SNPs.

### 5.2.3 Dataset

Genotyping and clinical information of Korean individuals was produced as a part of Korean Association Resource (KARE) project by Korean Centers for Disease Control and Prevention (Cho et al., 2009; Hong et al., 2012). The cohort study was examined for 8842 individuals at Ansan and Ansung area, aged 39 to 70. The genotyping was conducted using Affymetrix Genome-Wide Human SNP array 5.0. In the clinical information, we investigated the concentration of glucose, diagnosis, and drug treatments. If a person have been an experience to take a diagnosis for the diabetes, we decided the person have a diabetes. Also if plasma glucose is same to or over 126 (mg/dl) in no caloric intake or two-hour plasma glucose is same to or over 200 (mg/dl), then we considered the person a diabetes case. Conversely, the criterion for the normal controls are the plasma glucose with no calory intake is under 100, two-hour is under 140, and no experience for diabetes diagnosis or insulin intakes.

### Odds ratio

The odds ratio is used to measure a relative risk in a specific genotype comparing to another one. It was calculated as follows:

$$oddsratio = \frac{p_1(1-p_1)}{p_2(1-p_2)} \tag{5.8}$$

, where $p_1$ and $p_2$ are probabilities that an individual having the selected SNP rules exists in the disease group and normal group, respectively. If an odds ratio is greater than 1, the events is more likely to occur disease. That is, the odds ratio which is significantly higher than 1, means the higher-order SNPs are associated with disease.

The p-value is measured by random combination of SNPs. We generated 100,000 SNP rules randomly, and calculated odds ratio in each rule. Then we checked the probability that the odds ratio for the selected rules occurs by chance.

## 5.3 Results

### 5.3.1 Identifying interaction between features in simulation data

To verify our approach can find the interaction of features, we tested the method using simulation data. Suppose that the simulation data $X_i = (x_1, x_2, ..., x_{10}, class)(1 \leq i \leq 1,000)$ is composed of 10 attributes, $x_j \in 0, 1(1 \leq j \leq 10)$. By gibbs sampling, we generated the data with following conditions:

$$
\begin{aligned}
&P(x_1 = 0) = 0.6 \\
&P(x_2 = 0) = 0.6 \\
&P(class = 1|x_1 = 1 \wedge x_2 = 1) = 0.8 \\
&P(class = 1|x_1 = 1 \wedge x_2 = 0) = 0.3 \\
&P(class = 1|x_1 = 0 \wedge x_2 = 1) = 0.3 \\
&P(class = 1|x_1 = 0 \wedge x_2 = 0) = 0.2
\end{aligned}
\tag{5.9}
$$

$x_3$ and $x_4$ have same probabilities with $x_1$ and $x_2$, respectively and the others are randomly generated (uniformly distributed). Table 5.2 shows the finally selected interactions by our approach using the simulated data. As we expected, our method can find the informative interactions of features for the classification with around 0.70 classification performance by 10-fold cross validation. The set of $x_1$ and $x_2$ were selected as the most highly ranked interaction. Also, the pairs of $x_3$ and $x_4$ similarly had big weight values after the learning by our approach. Although a state of the art classifier, SVM, has a little higher accuracy (0.734), the algorithm does not provide which features are important for the classification. Moreover, it is impossible to detect the combinatorial effects among the genetic variants, but our method can it.

Table 5.2: Identified interaction in simulated data

| Interaction | Weight |
|---|---|
| $x_1$=1, $x_2$=1 | 0.91 |
| $x_1$=1, $x_2$=0 | -0.89 |
| $x_3$=1, $x_4$=1 | 0.81 |
| $x_3$=0, $x_4$=1 | -0.79 |
| $x_3$=1, $x_6$=1 | -0.79 |

### 5.3.2 Identifying higher-order SNP interactions in Korean population

Korean population might be specific associative characteristics to a disease. Since we confirmed that our proposed method would be adequate to be find combinatorial effects of SNPs in genome-wide association study, at last we searched the multiple SNP interaction in Korean population using our method.

For the preprocessing, we firstly carried out hardy-weinberg test (HWE), then filtered out uninformative SNPs (p value $< 0.000001$). Then we removed SNP attributes where minor allele frequency (MAF) is less than 0.01. Then for each SNP, the p-value was calculated based on a chi-square test. We also filtered out significant SNPs (p value $< 0.05$). After the preprocessing, the number of attributes was decreased to 6459.

The main purpose of our approach is to identify higher-order interaction of multiple SNPs, but it can be run as a classifier. Also, it is required to check the classification performance for selecting highly discriminative combination of SNPs. Table 5.3 shows classification accuracy in our method. Using 10-fold cross-validation, the classification accuracy was around 90% when we evaluated the performance along the chromosome length. We also carried out other classification algorithms using the same datasets and compared the accuracy (Table 5.3). Even though it had a little difference with the interaction length to be examined, we obtained better or competitive performance to the results of other general classifiers. Usually, tree-based classifiers can be used to know which factors affects to the classification. However, in the dataset, the tree-based methods were shown much lower classification accuracy, 61.11% with decision tree (ID3) and 70.66% with random forest, which is considered as a combination of decision rules in classification tree forms. The classification accuracies of other approaches were also about 73.59% with an instance based classifier (k-nearest neighbor, kNN) and 71.38% with logistic regression. Only RBF network and SVM achieved the similar accuracy to our method. However, these two algorithms do not provide which factors significantly affect to the classification. The results mean that, our approach can find higher-order interactions of SNPs by choosing the highly-weighted individuals from the learned models, along the chromosome lengths.

Table 5.3: Classification performance in KARE dataset

| Order ($l$) | Accuracy |
|---|---|
| $l$=2 | 91.20 |
| $l$=3 | 91.40 |
| $l$=4 | 89.16 |
| Decision Tree (ID3) | 61.11 |
| Decision Tree (C4.5) | 60.22 |
| Random Forest | 70.66 |
| kNN (k=10) | 73.59 |
| SVM | 94.81 |
| RBF Network | 92.83 |
| Simple Perceptron | 67.11 |
| Logistic Regression | 71.38 |

In each experiment, we selected top SNP combinations from the ranking of their weights, and subsequently, we evaluated significance of the interactions through the odds ratio and the chi-square test. Table 5.4 shows top 10 interactions as an experimental result with order 3, and Figure 5.2 show the interaction map. The highly positive value of the weight implies the interaction can be a big effect to T2D, and negative means it is affectable to be classified to the normal sample. The table presents that the positively weighted interactions all have the high (>1) odds ratio. Conversely, the interactions with negative weight values low (<1) odds ratios. That is, the results suggest that the positively-weighted interaction is able to be a candidate for the T2D risk factors. In addition, the interactions were significantly distinguishable between case and control data by a chi-square test. the p-values by a chi-square test were significantly low in the whole selected interactions.

Table 5.4: Highly ranked SNP interaction

| SNP interaction | | | weight | odd ratio | p-value (chi square test) |
|---|---|---|---|---|---|
| SNP_A-4196226 TT | SNP_A-2038226 CC | SNP_A-1861290 GG | -3.7 | 0.754 | 3.09e-4 |
| SNP_A-1963560 CC | SNP_A-4222651 TT | SNP_A-2032424 GG | -2.78 | 0.738 | 5.77e-5 |
| SNP_A-2144088 GG | SNP_A-2055282 CC | SNP_A-2293836 AA | 2.52 | 1.416 | 6.49e-6 |
| SNP_A-2182681 GG | SNP_A-4223259 TT | SNP_A-2033011 GG | 2.18 | 1.366 | 7.42e-5 |
| SNP_A-2269625 CC | SNP_A-2257007 CC | SNP_A-1788186 AA | 1.86 | 1.352 | 1.91e-4 |
| SNP_A-2178766 CC | SNP_A-1932380 CC | SNP_A-2224407 GG | -1.84 | 0.707 | 7.93e-6 |
| SNP_A-1842269 TT | SNP_A-2205381 CC | SNP_A-1982225 AA | 1.84 | 1.209 | 1.31e-2 |
| SNP_A-1829387 TT | SNP_A-1861385 CC | SNP_A-4283627 AA | 1.74 | 1.387 | 2.86e-5 |
| SNP_A-1802450 CC | SNP_A-2048106 GG | SNP_A-2033011 GG | 1.64 | 1.387 | 1.86e-5 |
| SNP_A-1869508 TT | SNP_A-1844690 AA | SNP_A-1884338 CC | 1.64 | 1.348 | 3.00e-4 |

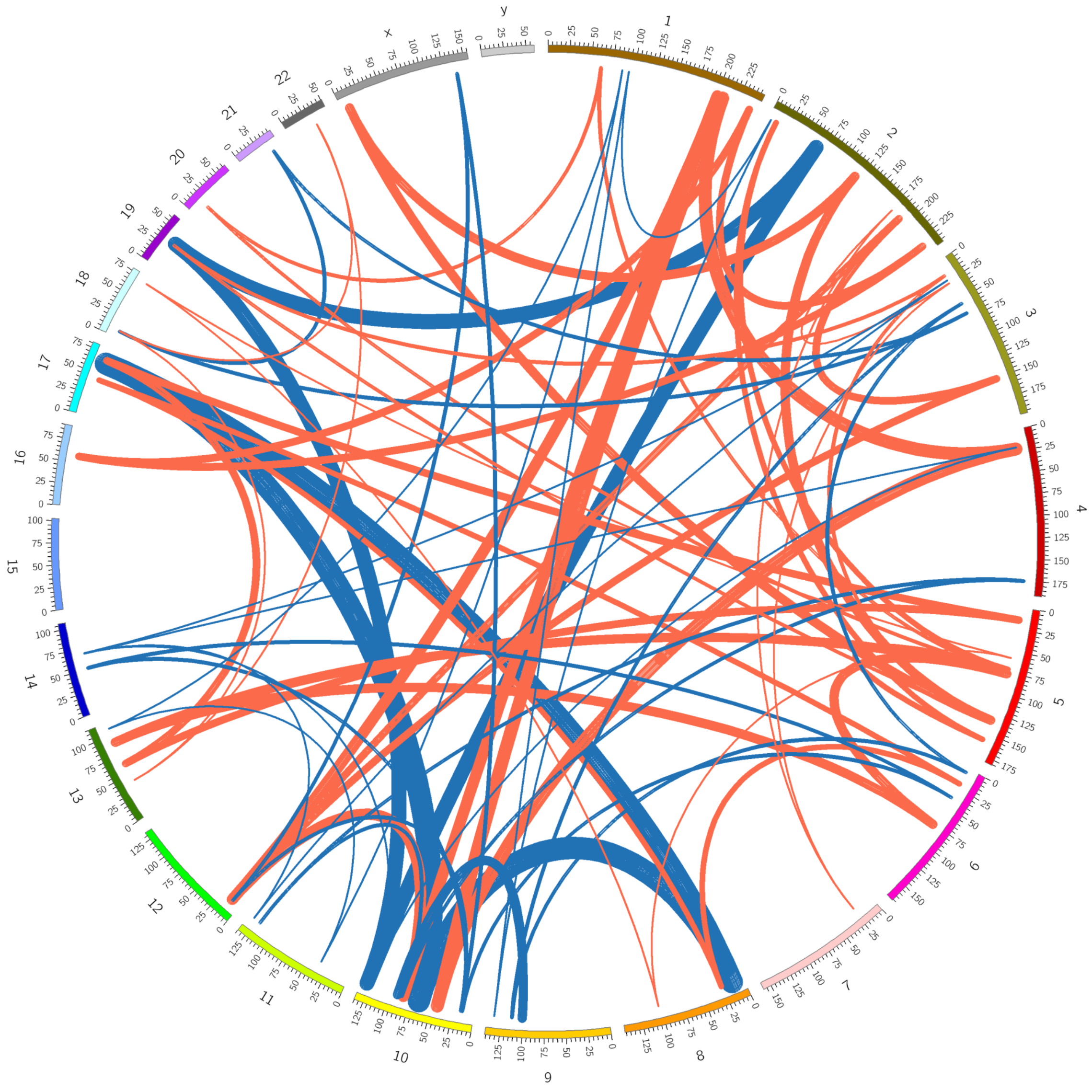


Figure 5.2: SNP interaction map order 3. The thickness of the lines means weights of the interactions. Blue and red colors mean negative and positive weights, respectively.

Interestingly, our results showed that the sequence variation could have much clear association with the higher-order interaction, even although it did not show the strong evidence in single-SNP analysis. Figure 5.3 shows the results for the top 5 ranked interactions. The p-values of the identified interactions were clearly lower than those in single variants within the interaction by our experiments. For an instance, the firstly ranked interaction, SNP_A-4196226, SNP_A-2038226 and SNP_A-1861290 did not show clear association with diabetes as a single variant. The p-values for the single SNP were 0.06, 0.04, 0.02, respectively. However the combination of these was definitely stronger effects to a disease with 3.09e-04 p-value.

For further validation, we randomly generated 1,000 interactions which consist of 3 SNPs and choose the interactions whose number of the matched to instances are more than 10. Then we measured their p-values by the chi-square test. Figure 5.4 shows the p-value comparison between top 100 interactions in our results and the randomly generated set. It shows that the interactions in our results are much more significant. When we carried out a t-test to clarify how these two sets are different, the p-value by the t-test was 9.79e-133.

## 5.4 Discussion

We presented a method to identify higher-order interaction of multiple variables. The study to identifying the higher-order interaction of genetic variants is necessary to find the multiple causal factors, contribute to complex diseases. Although the analysis of multiple factor interaction should be important in understanding complex traits, however, it is computationally infeasible to combinatorially explore all high-order interactions among the SNPs in a genome-wide association study. Previously several studies reported on findings of interactions among genes to be

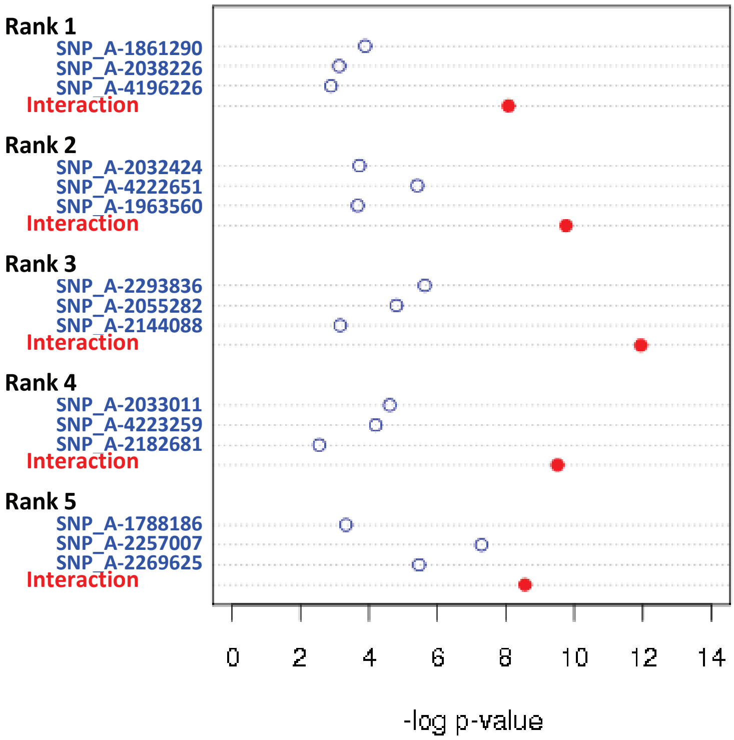


Figure 5.3: Dotchart for comparison between single variants and their interaction. Empty circle is for a single variant and filled is for the interaction.

important contributors to certain phenotypic variation. However, in addition to the variants of genes which directly changes protein function, the genetic alteration may be located in genomic or epigenetic regulatory regions. These can also affect to the gene regulation and abnormality in cellular processes.

We used evolutionary learning to search the combinatorial feature spaces. Generally evolutionary computation finds a good solution by a guidance from a fitness and genetic operators. Using the concepts, we could find a solution, coherent group

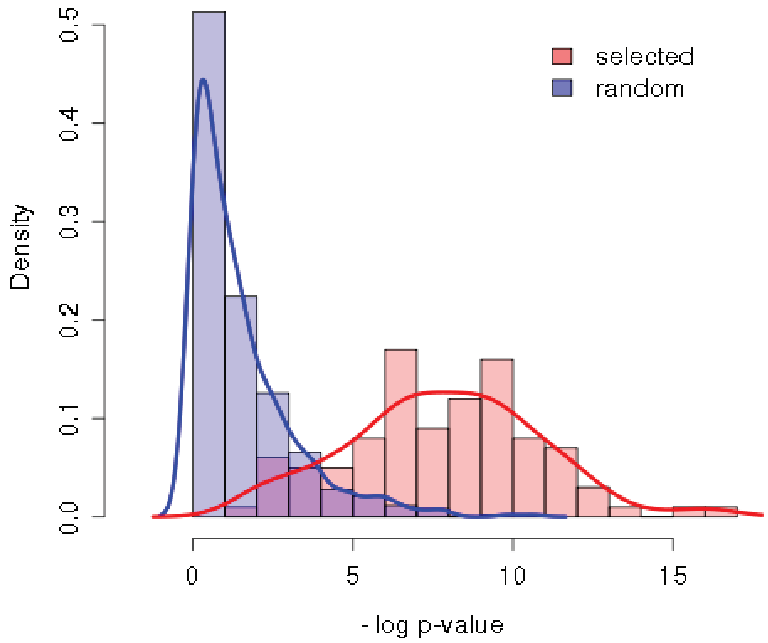


Figure 5.4: Comparison between the interaction by our approach and random selection. Red is a histogram for the identified interaction by our method, and blue is random selection.

of interrelated variants, associated to a disease effectively. When we examined every possible case, the search space is too big. For example, If the number of attributes is 6459 and the combinatorial order is increased from 2 to 5, the number of possible combinatorial cases are 2.09E07, 4.49E10, 7.25E13, and 9.35E16, respectively. However, we searched only cases less than 1.00E6 in every experiment and could find reasonable high order interaction associated to disease. Our genetic association studies for complex traits can be applied to a systems genetics studies integrated with other information, such as environmental factors, copy number variation, clinical information, and so on. The systems genetics approach helps to yield a detailed map of genetic and other variants, including environments, associated with phenotypes. Our proposed methods can easily add these factors in steps to generate individuals, and find their effects to a disease. Also, the evolutionary learning in our approach make it possible to control the large datasets with a explorative search space. so a number of factors can be supplied in the consistent algorithm.

In our experiments, we did not reflect biological knowledge or genetic relationships. Depending on a experimental purpose, these information can be reflected in the process on generation of individuals or in the fitness function. Or it is also possible to construct a model with genetic relationships by measuring linkage blocks or conducting a transmission disequilibrium test from datasets.

In addition, the analysis of the interaction accompanies several issues including information loss with missing values. But our approach does not require imputation of the missing values, and it can be run by denoting these missing values as don't care symbols or mismatched symbol.

Sometimes, a sampling approach is an efficient method to find an optimal solution in a large datasets. However the datasets would be too sparse, especially in case of higher-order combination of variants. So we should randomly generate some of the individuals, instead of sampling from training datasets. In addition, if we want to search the interactions between just two variables, it might be not necessary to use crossover or mutation. It could be possible to find the fittest one, by random sampling in the reproduction processes. But in the interactions of multiple variables, it would be efficient to use these operators.

In each experiment, the chromosome length was constant. If the experiments are carried out to identify interactions with a variety length at one time, the individuals with a small length are more likely to be matched to a datasets, so it can be much bigger weight values. However, our approaches can be easily expanded to a method for identifying the interactions of variable length. One way is to normalize the fitness value by the chromosome length. Then we can find the interactions of various orders, resulted from individuals with diverse lengthes. Another way is to learn from lower to higher order by turns, and then to re-learn and classify based on the finally survived individuals in each step.

By the characteristics of evolutionary learning, our results would not be global optima. But it is definitely valuable. Our purpose is not to find one optimal coherent variant set associated to a disease. Also it might be impossible to be expressively provided that the complex traits are caused only by a little number of factors. The reason of the disease occurrence is not simple. Therefore, we detect interactions which may be local optima and provide the candidates to help to find sets of the risk factors.

Recent advances in high-throughput sequencing provide a variety of datasets. The sequencing datasets may shed light for a new finding in the GWAS, and whole-genome or whole exome sequencing has been used to search the genetic cause of diseases. Despite of the considerable progress in the sequencing technologies and their analysis strategies, the common variations identified by GWAS account for only a small fraction of disease heritability and are unlikely to explain the majority of phenotypic variations of common diseases. Our approach can be usefully applicable to the sequencing datasets. The sequencing technologies have detected millions of novel variants. Although big size of dataset by lots of reads and variants is another challengeable problem, our approach can be a method to solve the problem by effectively searching the combinatorial feature space based on the evolutionary learning. It can be a effective method to systematically control exploration of a lot of variants provided by next generation sequencing technologies for GWAS. Also, the sequence datasets have a large proportion of missing data, but our method can be resistable.

Our approach suggests the analysis of GWAS datasets offers a useful strategy for identifying causal genes and potential candidates in human diseases. Study for interaction of the genes or genomic regions would help to elucidate mechanism of the complex traits and to control and treat disease. Some of our results do not show clear relationships and some of these may be still biologically questionable, why the combination is highly weighted and how there play a role in disease. For the much clear understanding, relevant functional studies should be carried out. Moreover, by applying phased haplotype information, we will detect much relevant sets for variants (Tewhey et al., 2011).

# Chapter 6

# Identifying DNA methylation modules by probabilistic evolutionary learning

## 6.1 Background

Genomics mainly aims to find genetically associated markers with a phenotype. Based on DNA sequences, the researchers search causal effects to biological processes including gene regulatory mechanism and disease. Although several risk factors were identified by the association studies, the genetic variants do not fully explain the abnormal regulation, since the biological regulatory mechanism can be affected by many other factors, as well as DNA sequence modification (Jones and Baylin, 2007; Sadikovic et al., 2008; Handel et al., 2010; Sandoval and Esteller, 2012).

Epigenomics refers to a study for regulation of various genomic functions that are controlled by another partially stable modification, not DNA sequence variants (Bonetta, 2008). Among these, DNA methylation, which typically occurs at CpG dinucleotide by DNA methyltransferase (DNMT) enzyme, is a crucial epigenetic regulatory mechanism in cellular processes. The DNA methylation of CpG site mostly cause silence of the downstream gene, so the enrichment of the differentially methylated DNA fractions can contribute to specific abnormalities, including complex diseases (Robertson, 2005; Portela and Esteller, 2010; Jones, 2012). Especially, with an advent of microarray and next generation sequencing (NGS) technology, many researchers have carried out genome-wide DNA methylation profiling studies (Laird, 2010; Hill et al., 2011; Rhee et al., 2013), and the genome-wide studies have reported that lots of genomic regions are differentially methylated in normal and abnormal cells (Cheung et al., 2010; Toperoff et al., 2012; Walker et al., 2011).

However, it is well-known that a complex disease is generally caused by combinatorial dis-regulatory effects of multiple genes (Hirschhorn and Daly, 2005; Janssens and van Duijn, 2008; Kiezun et al., 2012). That is, the errors of biological processes is not caused by alteration of an individual methylation level. Recently, Easwaran et al suggested a concept for DNA hypermethylation modules which preferentially target important developmental regulators in embryonic stem cells (Easwaran et al., 2012). They found the set of genes by the DNA methylation would be contribute to stem-like state of cancer. Horvath et al. studied aging effects of DNA methylation and showed there exist co-methylated modules related to aging in human brain and blood tissue (Horvath et al., 2012).

Here, we identity combinatorial modules of DNA methylation sites associated to human disease by an evolutionary learning approach. The evolutionary algorithms can approximate solution well in lots of problems (Kumar et al., 2010; Deb and Datta, 2010; Joung et al., 2012; Wang et al., 2013). It generates new population through iterative updates and selection by a guided search process in a feature space. We utilized an estimation of distribution algorithm (EDA)-based learning approach for identifying combination of cancer-related DNA methylation sites. In the EDA algorithm, the population is evolved according by probabilistic distribution in the selected individuals without conventional genetic operators such as crossover and mutation. It has been known that EDA efficiently and effectively provide answers in combinatorial optimization problems (Chen et al., 2009; Zhou et al., 2009; Shim et al., 2013; Ceberio et al., 2013). The EDA has been previously applied in several biological research, and it has offered promising results for complex problems, in where other methods fail to find good solution (Pal et al., 2006; Santana et al., 2010; Shelke et al., 2013).

In this study, we investigated DNA methylation modules relevant to cancer, using the DNA methylation profiling datasets produced by microarray- and sequencing-based approaches. The experimental results show that our method can find the DNA methylation modules well related to cancer.

## 6.2 Methods

### 6.2.1 Evolutionary learning procedure to identify a set of DNA methylation sites associated to disease

EDAs evolve a population to find optimal solution probabilistically. The initial population is composed by constructing individuals at random. The individual represents higher order interaction of the methylated sites. The population size $m$ is decided empirically and the initial weight $w_j$ of the individual $j$ $(0 < j < m)$ is randomly assigned with a small value (-1 $< w_j <$ 1).

In the evolutionary process, each individual is evaluated how the interaction is discriminative for the datasets. Then, the better individuals are selected and the dependency tree fitted to the selected individuals, is build. New individuals of the next generation are generated using the probability distribution within the tree structure, and replace the previous individuals. The overall procedure follows:

1. Set $g \leftarrow 0$
2. Initialize population $X(g)$ by random generation
3. Evaluate individuals in $X(g)$
4. Select a set of individuals by tournament selection from $X(g)$
5. Construct a dependency tree $G(g)$ by measuring Kullback-Leibler divergence between variables
6. Parameter learning using probability distribution of the selected set
7. Generate a new individuals by sampling with joint distribution from the $G(g)$, and create new population $X(g+1)$
8. Set $g \leftarrow g+1$
9. If the termination criterion is not met, go to 3

More details for steps 3 and 5 are explained in following sections.

### 6.2.2 Learning dependency graph

The dependency tree is built from the selected individuals by searching conditional dependencies between random variables. Then the model is optimized by a series of incremental updates (Pelikan, 2006; Pelikan et al., 2007). More details follow:

Suppose that $X$ is population and $X = \{X_1, X_2, ..., X_n\}$ is presented as a vector of variables with $n$ features, that is, DNA methylation sites. The probability distribution is represented by a joint probability $P(X_1, X_2, ..., X_n)$ as to:

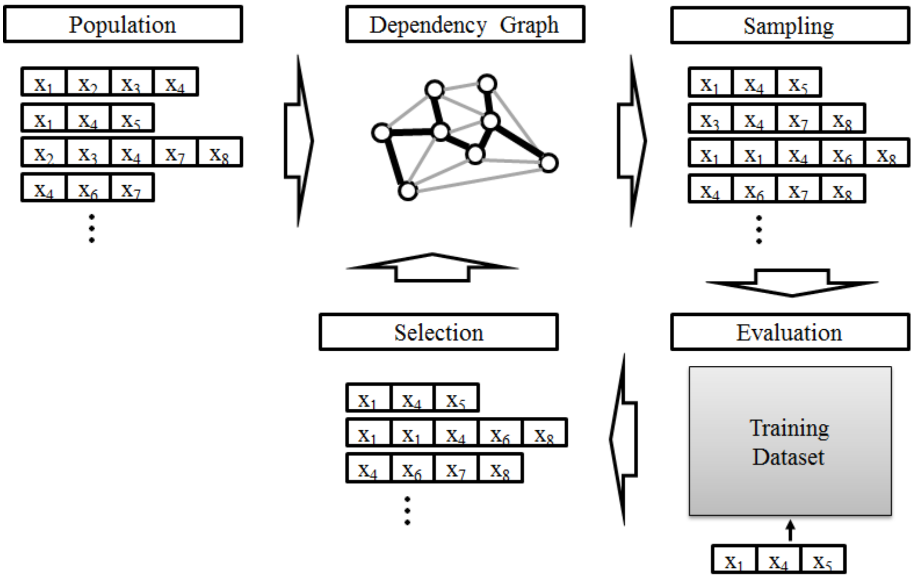


Figure 6.1: Schematic overview for probabilistic evolutionary learning to identify DNA methylation module, Iterative evolutionary learning.

$$\begin{aligned} P(X) &= P(X_1, X_2, ..., X_n) \\ &= P(X_1|X_2, ..., X_n)P(X_2|X_3, ..., X_n)....P(X_{n-1}|X_n)P(X_n). \end{aligned} \tag{6.1}$$

However, it is hard to measure all the joint probabilities exactly when $n$, the number of variables, is large. Thus it needs to approximate the probability distribution. For the purpose, in this study, we used a dependency tree, and the distribution is approximated as follows:

$$P(X_1, X_2, ..., X_n) = P(X_r) \prod_{i \neq r} P(X_i|X_{pa(i)}), \tag{6.2}$$

where $X_1, X_2, ..., X_n$ are random variables, $r$ is an index of root node, and $pa(i)$ denote the index of parent node of $X_i$.The tree structure is built by searching based on Kullback-Leibler divergence between two random variables. The dependency graph is optimally constructed in a direction to maximize total mutual information as follows:

$$argmax_{r,pa}\prod_{i\neq r} I((X_i)pa(i)) \tag{6.3}$$

$$\begin{aligned} &I((X_i)pa(i)) = \\ &\sum_x\sum_y P(X_i = x, X_{pa(i)} = y) log\frac{P(X_i = x, X_{pa(i)} = y)}{P(X_i = x)P(X_{pa(i)} = y)} \end{aligned} \tag{6.4}$$

The complete graph $G$ searches the maximum spanning tree, and then the best dependency tree is constructed.

For parameter learning, the most likely values are calculated from the frequencies in the selected individuals. That is, the model parameters are represented as a marginal probabilities in a root node and conditional probabilities in the other nodes. The marginal probabilities in root nodes and the conditional probabilities in child nodes are calculated as:

$$P(X_r = x) = \frac{m(X_r = x)}{N}, \tag{6.5}$$

$$P(X_i|X_{pa(i)}) = \frac{m(X_r = x)m(X_i = x, X_{pa(i)} = y)}{m(X_{pa(i)})}. \tag{6.6}$$

### 6.2.3 Fitness evaluation in population

Each individual has a fitness value which means how informative the chromosome is to classify the samples. That is, the fitness for individuals are evaluated by measure the classification accuracy for interaction of the features. To determine and update the fitness for each individual, it is possible to use any classification algorithm. But we introduce a gradient descendant rule for training data $\mathbf{D}$ as follows:

$$w_i = w_i + \eta(t_j - f(\mathbf{D}_j))m_{ji}, \tag{6.7}$$

where $w_i$ is a weight value for $i$-th feature and $t_j$ is a target class in the $j$-th training instance $\mathbf{D}_j$. $\eta$ is a learning rate and $m_{ji}$ is a value of the $i$-th attribute in the $j$-th instance. $f(\mathbf{D}_j)$ is a predicted output value of the $j$-th training instance by our model and determined as follows:

$$f(\mathbf{D}_j) = \begin{cases} 1, & \text{if } \sum_{i=0}^{n} w_i \cdot m_{ji} > 0 \\ -1, & \text{else} \end{cases} \tag{6.8}$$

The difference between the predictions and the target values specified in the training sequence is used to represent the error of the current weight vector. The target function is optimized to minimize the classification error. The weight values are evaluated against a sequence of training samples and are updated to improve the classification accuracy. The weight update processes are repeated until it would be converged after the number of epoch.

Using the learning scheme, we find most informative individuals for classification, that is, their absolute value of their weights is large. In addition, since our purpose is to identify a DNA methylation module, it might be necessary to find it if the number of the used feature is small. Finally, the fitness function for the $k$-th individual $X^k$, $Fitness(X^k)$ is defined as follows:

$$Fitness(X^k) = Acc(X^k) - Order(X^k), \tag{6.9}$$

where $Acc(X^k)$ is classification accuracy for training datasets and $Order(X^k)$ denotes the number of methylation sites which is selected in the individual $X^k$.

### 6.2.4 Dataset

The high-throughput DNA methylation profiling of large genomic regions could be produced by both microarray and NGS technologies. We applied our approach to these two types of datasets. The microarray data was generated by Illumina Infinium 27k Human DNAmethylation BeadChip in 1,475 samples, for surveying of genome-wide DNA methylation profiles in breast cancer and normal samples (Zhuang et al., 2012). Sequence-based datasets were produced by MethylCap-seq in matched normal and colon cancer samples and collected at GSE39068 (Simmer et al., 2012). The normalization and preprocessing was carried out using the same approaches to Simmer's works (Simmer et al., 2012).

## 6.3 Results

### 6.3.1 DNA methylation modules associated to breast cancer

This analysis has been carried out based on DNA methylation profiling datasets which experimentally measured the methylation status using DNAMethylation BeadChip (Zhuang et al., 2012). We extracted data for DNA methylation profiles on chromosome 17 from breast cancer and normal samples, and applied our method to the dataset. Figure 6.2 shows learning curves in the evolutionary process. The fitness value is improved when the number of generation is increased. Since we introduced a term for the number of the methylation sites to find a individual with the shorter length, the number of orders were decreased at the learning process (Figure 6.2). After convergence, 6 sites were selected for the discrimination, and these 6 sites are related to genes, KIAA1267, CD79B, ALOX12, TMEM98, KRT19 and FOXJ1 (Table 6.1).

ALOX12 have a role in growth of breast cancer and its inhibition may be a

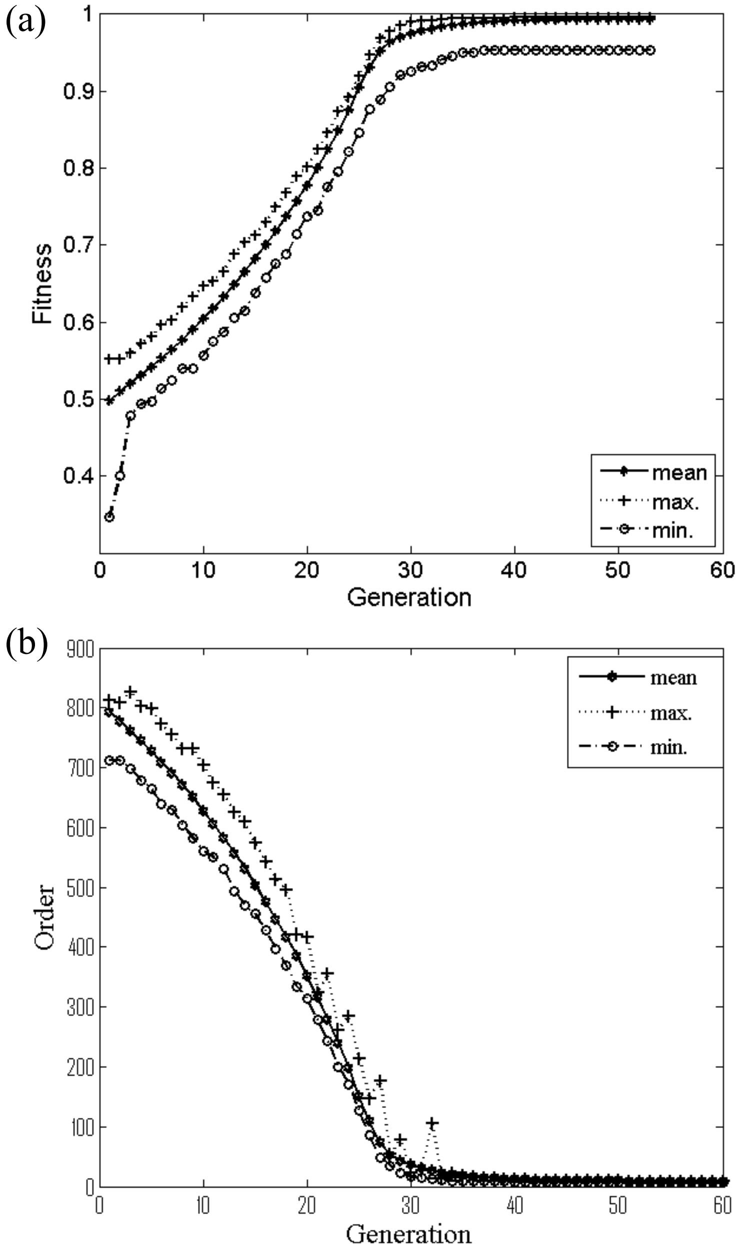


Figure 6.2: Learning curve using breast cancer datasets. x-axis is the number of generation and y-axis is (a) fitness values and (b) the number of orders.

Table 6.1: Finally selected methylation sites

| ID | Position | Gene | CGI location |
|---|---|---|---|
| cg02301815 | 41605268 | KIAA1267 | 41605074-41605445 |
| cg07973967 | 59363339 | CD79B | 25467633-25468370 |
| cg08946332 | 6840612 | ALOX12 | 6839463-6841283 |
| cg11833861 | 28279748 | TMEM98 | 28278827-28279833 |
| cg16585619 | 36938776 | KRT19 | NaN |
| cg24164563 | 71647990 | FOXJ1 | 71647419-71649480 |

strategy for inhibiting tumor growth (kumar Singh et al., 2012), the gene can be used as a serum marker for breast cancer (Singh et al., 2011). It is not clearly known how the ALOX12 methylation directly affects to breast cancer. However, it has been reported that hypermethylation of ALOX12 can be associated to cancer (Tan et al., 2009; Alvarez et al., 2010; Ammerpohl et al., 2012; Ohgami et al., 2012). Actually, the ALOX-12 gene is closely related to apoptosis, and the problem of the expression by the DNA methylation can cause a malfunction of the cell death (Ding et al., 1999; Pidgeon et al., 2002, 2003). Therefore, it might be reasonable that the change of methylation in the gene linked to most cancer, including breast tumor. KRT19 is a well-known marker for breast cancer patients (Ring et al., 2004; Lacroix, 2006), and KRT19 promoter is abberently methylated in cancer cell lines (Morris et al., 2008). Also, it has been reported that there exist the relationships between expression of CD79B and breast cancer (Ellsworth et al., 2008; Prat et al., 2010). FOXJ1, a member of the forkhead box (FOX) family, may function as a tumor suppressor gene in breast cancer (Jackson et al., 2010). FOXJ1 is hypermethylated and silenced in breast cancer cell lines (Demircan et al., 2009). TMEM98 is one of transmembrane

Table 6.2: Classification performance only using the 6 selected sites

| Algorithm | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Logistic Regression | 0.939 | 0.987 | 0.762 |
| SVM | 0.929 | 0.941 | 0.857 |
| Decision Tree | 0.939 | 0.952 | 0.867 |
| Naive Bayes | 0.919 | 0.951 | 0.765 |

proteins. Recently, Grimm et al. investigated the transmembrane proteins specific for cancer cells. The transmembrane protein can be a target for antibodies and be a biomarker for tumor diagnosis, prognosis, and treatment (Grimm et al., 2011). The function of KIAA1267 is not clearly known yet. But the gene encodes KAT8 regulatory NSL complex subunit 1, and the KAT8 regulates p53, a tumor suppressor gene (Li et al., 2009; Zhang et al., 2013). It imply the KIAA1267 can has a role in breast cancer.

Using the 6 sites, we tested classification performance using general machine learning algorithms (Table 6.2). To verify our method identified informative sites, we carried out classification only using the selected features. Table 6.2 shows the classification accuracy, sensitivity and specificity. Regardless the classification algorithms, it could be well-classified. For further verification, we randomly extracted the sites repeatedly (100 times), then measured the classification performance in each dataset. Figure 6.3 shows that the results of our method were higher than others, regardless of the number of the randomly selected sites.

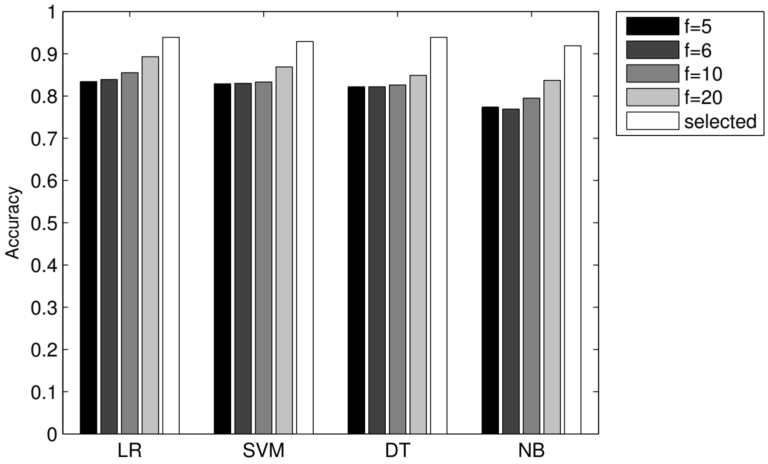


Figure 6.3: Classification accuracy using randomly selected sites. f is the number of the randomly selected sites, and white bar, marked as selected, is the results using only the 6 selected sites by our method. The results for the random datasets show averages of 100 times repeated experiments. LR: logistic regression, SVM: support vector machine, DT: decision tree, NB: naive Bayes.

### 6.3.2 Modules associated to colorectal cancer using high-throughput sequencing data

Recently, DNA methylation profiles could be measured by high-throughput sequencing technologies. We applied our method to the sequencing-based methylation profiling datasets produced by Simmer et al. (Simmer et al., 2012).

Figure 6.4 depicts improvement of the fitness in iterative learning procedures using these datasets. Among 10,393 genomic regions on chromosome 17 for the experiment, 348 regions were selected to discriminate the ovarian cancer and normal

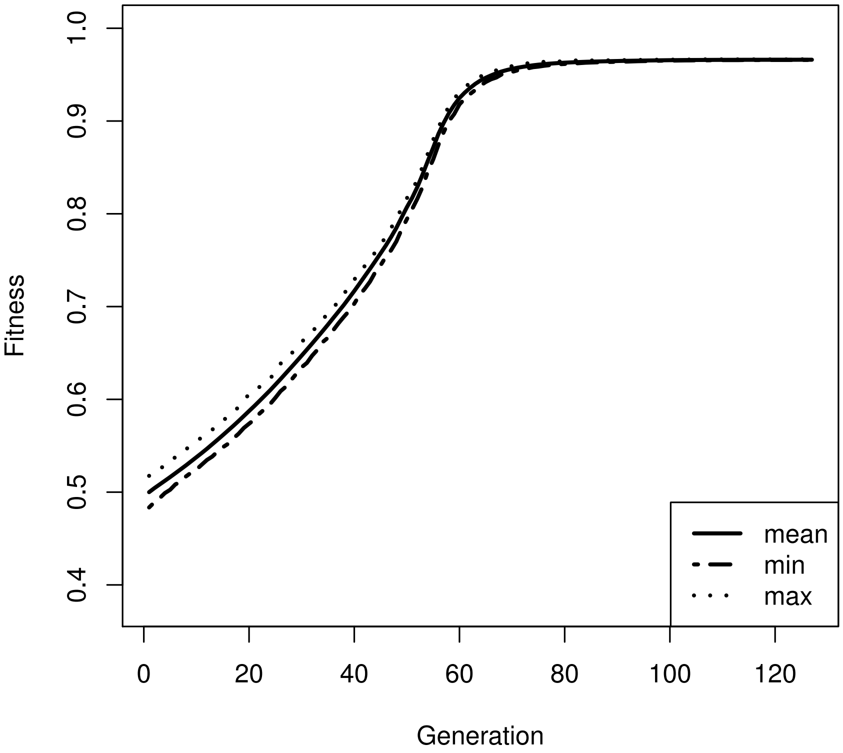


Figure 6.4: Learning curve using in colon cancer datasets. x-axis is the number of generation and y-axis is (a) fitness values.

samples after a convergence. Table 6.3 shows performance by classification algorithms using the 348 regions from the sequencing-based colorectal cancer datasets.

We annotated the selected regions using GPAT (Krebs et al., 2008) and investigated which genes were located closely on the selected regions. We accomplished gene set enrichment analysis (GSEA) with KEGG pathway using the genes whose

Table 6.3: Classification performance only using the 348 selected sites in colorectal cancer data

| Algorithm | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Logistic Regression | 0.900 | 0.920 | 0.880 |
| SVM | 0.940 | 0.960 | 0.920 |
| Decision Tree | 0.640 | 0.680 | 0.600 |
| Naive Bayes | 0.900 | 0.920 | 0.880 |

transcription start sites are located within 5000bp from the selected genomic regions. The GSEA was carried out using MSigDB (Subramanian et al., 2005; Liberzon et al., 2011). Table 6.4 summarizes the significantly enriched pathways with low p-values and shows that most of these are closely associated with cancer-related networks. Table 6.5 show the genes commonly enriched in the pathways. Note that the enriched signalling pathways were related to colorectal cancer. In colon cancer, the roles of wnt signalling pathway and MAPK signalling pathway have been very well-known (Jansson et al., 2005; Segditsas and Tomlinson, 2006; Fang and Richardson, 2005; Slattery et al., 2012). The genetic mutation affecting the pathway components and the alteration of their expression can enhance tumorigenicity in cancer cells. Also, neurotrophin signalling pathway could be related to growth of colorectal cancer cells (Akil et al., 2011) and chemokine signalling pathway suppresses colon cancer metastasis (Kitamura et al., 2010; Chen et al., 2012). Phosphatidylinositol signalling pathway plays an important role in the growth, survival and metabolism of cancer cells, and targeting this pathway has potential to lead to treatments for the colon cancer (Parsons et al., 2005; Yuan and Cantley, 2008). VEGF and ErbB can be valid therapeutic targets for patients with colon cancer (Ellis and Hicklin, 2008; Winder and Lenz, 2010; Roskoski Jr, 2004; Spano et al., 2005).

For further validation, we compared the results with ChIP-seq profiles of H3K4me3 and H3K27me3 at ENCODE project (Dunham et al., 2012). When we examined the selected sites on promoter regions, many of those were overlapped with the H3K4me3 and H3K27me3 binding sites with p-values of 1.86E-11 and 1.94E-05, respectively. The p-value for the regions overlapped with both of the two histone marks was 1.08E-05. The binding regions of the histone modification, called bivalent regions, were associated to cancer formation by abberant DNA methylation which leads to be silencing of regulators (Young et al., 2011; Chapman-Rothe et al., 2012). Since it is possible that DNA methylation are associated to bivalent regions in cancer, our studies would be help to understand the relationship between DNA methylation and chromatin signatures (McGarvey et al., 2008; Sharma et al., 2010; Balasubramanian et al., 2012; Reddington et al., 2013). Also it would help to investigate effects on cancer progression and possibilities for epigenomic treatments in cancer (Rodriguez et al., 2008; Mayor et al., 2011).

## 6.4 Discussion

DNA methylation can be also strongly associated with the complex diseases. It has been known that lots of genes are differentially methylated in various cancers or diseases. In this study, we presented a method to identify combinatorial effects of DNA methylation at multiple sites. From a systematic perspective, the relationship between DNA methylation regions and a specific disease is learned by the presented probabilistic evolutionary learning. The fitness value of a DNA methylation module measure the level of their responses to the disease. In computational view, our method can solve large scale problems by identifying modules with both compactness and high coverage of disease related genes. If the number of attributes is $n$, the number of possible cases is same to the number of elements in power set, $2^n$. Thus, the number of cases is exponentially increased according to the number of attributes. For example, if $n$=100, the number of cases is 1.27E30 and if $n$=1000, then the number is 1.07E301. However, Our method can find candidates in reasonable search in the problem space. In our every experiment, we found the candidate solution by searching less than 1.00E6 cases.

Applying our method to breast cancer and colorectal cancer data produced by high-throughput technologies, we detected the cancer-related modules confirmed by literatures and functional enrichment analysis. Interestingly we observed that the selected regions were located around genes which are enriched in cancer-related gene set categories significantly, and it provides evidence that the identified module in our study is biologically meaningful.

The studies on DNA methylation are likely to elucidate on the process of tumorigenesis as well as identify biomarkers. Our approaches which assist in the identification of multiple DNA methylation sites that have the potential to be epigenetically regulated might provide a useful strategy to detect epigenetic association related to a disease. The systematic identification of the disease-related genes and modules can provide insights into mechanisms underlying complex diseases and help efficient therapies or effective drugs.

By applying our method to microarray- and NGS-based data, we showed that it is applicable to a variety of data types and various disease contexts. Moreover, recent studies suggest that there exists a complex relationship between genetic variation, DNA methylation and so on. Systems genetic/epigenetics approaches are required for examining relationships among these. Although our framework is based on DNA methylation profiling datasets, it can be attempted to identify the combinatorial association for various factors including gene expression levels, microRNAs, copy number variation, genetic variations, and environmental factors. The integration of a variety of data would provide the basis for new hypothesis and experimental approaches in a model of complex disease.

In summary, we presented a method for searching the higher-order interaction of DNA methylation sites by a probabilistic evolutionary learning method. Using the approach, we also examined the potential for combined effects of various sites on genome. The results suggest that the alteration of DNA methylation at multiple sites affects on cancer. Similar to genome-wide association studies, our approach provides an opportunity to capture the complex and multifactorial relationship between the DNA methylation sites and to find new factors for future study. Therefore, our approach would be a way to facilitate a comprehensive analysis of genome-wide DNA methylation datasets and the interpretation for the effects of DNA methylation on multiple sites.

Table 6.4: Gene-set enrichment analysis annotated by promoter information using the 348 selected sites in colorectal cancer data

| Gene set | p-value | FDR q-value |
|---|---|---|
| Non-small cell lung cancer | 2.61E-05 | 4.25E-03 |
| Glioma | 4.56E-05 | 4.25E-03 |
| Neurotrophin signaling pathway | 3.25E-04 | 1.85E-02 |
| Pathways in cancer | 3.99E-04 | 1.85E-02 |
| Wnt signaling pathway | 5.52E-04 | 2.05E-02 |
| Aldosterone-regulated sodium reabsorption | 9.09E-04 | 2.22E-02 |
| Endocytosis | 9.62E-04 | 2.22E-02 |
| Vasopressin-regulated water reabsorption | 9.97E-04 | 2.22E-02 |
| Chemokine signaling pathway | 1.07E-03 | 2.22E-02 |
| Focal adhesion | 1.26E-03 | 2.34E-02 |
| Endometrial cancer | 1.39E-03 | 2.35E-02 |
| Basal cell carcinoma | 1.55E-03 | 2.41E-02 |
| Colorectal cancer | 1.97E-03 | 2.73E-02 |
| Pancreatic cancer | 2.50E-03 | 2.73E-02 |
| Melanoma | 2.57E-03 | 2.73E-02 |
| Chronic myeloid leukemia | 2.72E-03 | 2.73E-02 |
| Cytokine-cytokine receptor interaction | 2.82E-03 | 2.73E-02 |
| MAPK signaling pathway | 2.82E-03 | 2.73E-02 |
| Phosphatidylinositol signaling system | 2.94E-03 | 2.73E-02 |
| VEGF signaling pathway | 2.94E-03 | 2.73E-02 |
| Fc epsilon RI signaling pathway | 3.17E-03 | 2.81E-02 |
| Small cell lung cancer | 3.58E-03 | 2.98E-02 |
| ErbB signaling pathway | 3.83E-03 | 2.98E-02 |
| Apoptosis | 3.92E-03 | 2.98E-02 |
| Prostate cancer | 4.01E-03 | 2.98E-02 |

Table 6.5: Genes enriched in pathway analysis

| Gene Symbol | Description |
|---|---|
| TP53 | tumor protein p53 |
| PIK3R5 | phosphoinositide-3-kinase, regulatory subunit 5, p101 |
| PRKCA | protein kinase C, alpha |
| ARHGDIA | Rho GDP dissociation inhibitor (GDI) alpha |
| FZD2 | frizzled homolog 2 (Drosophila) |
| RABEP1 | rabaptin, RAB GTPase binding effector protein 1 |
| CCL16 | chemokine (C-C motif) ligand 16 |
| CXCL16 | chemokine (C-X-C motif) ligand 16 |
| CSF3 | colony stimulating factor 3 (granulocyte) |
| DUSP3 | dual specificity phosphatase 3 |
| ARSG | arylsulfatase G |

# Chapter 7

# Conclusion

Recently, explosive growth in data produced from various areas is continuously increasing. Intuitively the large amount of stored data contains valuable hidden knowledge, such that it could be used to improve the decision making process of an organization. There exists a clear need for the systematic methods for extracting the valuable knowledge from real-world datasets. This need has led to the emergence of a field called data mining and knowledge discovery. In order to extract or mine the knowledge or pattern of interest from data, intelligent mining tools are applied. The examples are association rule mining, clustering, classification, and so on.

Data collected from various biological domains is also becoming increasingly high in recent time. In particular, the large repositories of genome-wide measurement data pose the research question of how to retrieve valuable knowledge. In this dissertation, we proposed methods to identify higher-order interaction in genomic/epigenomic studies. We developed machine learning methods with evolutionary computation for extracting valuable information from large, high-dimensional data sets.

Statistical learning and evolutionary computation can be an way to mine the meaningful information from the biological big data. Especially, evolutionary computation has advantages to deal with a huge amount of the heterogeneous biological data. It appears to be more efficient in finding acceptable solutions than other random or semi-random search methods. Moreover, the approaches can be easily run in parallel, and allow groups of processor to be utilized for a search in the big data.

Furthermore, it might be helpful to exploit additional data sets even if they are only partially relevant for the data set of interest. For example, to further comprehensive understanding complex disease, it needs for integrative studies of various genomic and epigenomic datasets with environmental factors (Aschard et al., 2012). One advantage of our evolutionary machine learning approach is that it can easily extend and generalize the learning paradigms for multiple views of datasets. By systematically linking the various data sets, we would increases a chance to clarify biological knowledge and novel possibilities for biological results.

# Bibliography

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

Akil, H., Perraud, A., Mélin, C., Jauberteau, M.-O., and Mathonnet, M. (2011). Fine-tuning roles of endogenous brain-derived neurotrophic factor, trkb and sortilin in colorectal cancer cell survival. *PloS One*, 6(9):e25097.

Alvarez, S., Suela, J., Valencia, A., Fernández, A., Wunderlich, M., Agirre, X., Prósper, F., Martín-Subero, J. I., Maiques, A., Acquadro, F., et al. (2010). Dna methylation profiles and their relationship with cytogenetic status in adult acute myeloid leukemia. *PloS One*, 5(8):e12197.

Ammerpohl, O., Pratschke, J., Schafmayer, C., Haake, A., Faber, W., von Kampen, O., Brosch, M., Sipos, B., von Schönfels, W., Balschun, K., et al. (2012). Distinct dna methylation patterns in cirrhotic liver and hepatocellular carcinoma. *Int. J. Cancer*, 130(6):1319–1328.

Aschard, H., Lutz, S., Maus, B., Duell, E. J., Fingerlin, T. E., Chatterjee, N., Kraft, P., and Van Steen, K. (2012). Challenges and opportunities in genome-wide environmental interaction (gwei) studies. *Hum. Genet.*, 131(10):1591–1613.

Bach, F. R. and Jordan, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48.

Bailey, T. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings International Conference on Intelligent Systems for Molecular Biology; ISMB.*, volume 2, page 28.

Balasubramanian, D., Akhtar-Zaidi, B., Song, L., Bartels, C. F., Veigl, M., Beard, L., Myeroff, L., Guda, K., Lutterbaugh, J., Willis, J., et al. (2012). H3k4me3 inversely correlates with dna methylation at a large class of non-cpg-island-containing start sites. *Genome Med.*, 4(5):47.

Banerjee, N. and Zhang, M. Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, 31(23):7024–7031.

Baylin, S. B. and Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer*, 11(10):726–734.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., Pritchard, J. K., et al. (2011). Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol.*, 12(1):R10.

Bernado-Mansilla, E. and Garrell, J. (2003). Accuracy-based learning classifier systems: models, analysis and applications to classification tasks. *Evol. Comput.*, 11(3):209–238.

Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128(4):669–681.

Bloushtain-Qimron, N., Yao, J., Snyder, E. L., Shipitsin, M., Campbell, L. L., Mani, S. A., Hu, M., Chen, H., Ustyansky, V., Antosiewicz, J. E., et al. (2008). Cell type-specific dna methylation patterns in the human breast. *Proc. Natl. Acad. Sci. U. S. A.*, 105(37):14076–14081.

Bonetta, L. (2008). Epigenomics: Detailed analysis. *Nature*, 454:795–98.

Brāzma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8(11):1202–1215.

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D., and Scandura, J. M. (2011). Dna methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One*, 6(1):e14524.

Brinkman, A. B., Simmer, F., Ma, K., Kaan, A., Zhu, J., and Stunnenberg, H. G. (2010). Whole-genome dna methylation profiling using methylcap-seq. *Methods*, 52(3):232–236.

Brown, R. L., Reinke, L. M., Damerow, M. S., Perez, D., Chodosh, L. A., Yang, J., and Cheng, C. (2011). Cd44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J. Clin. Invest.*, 121(3):1064.

Bush, W. and Moore, J. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, 8:12.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.*, 27(2):167–174.

Ceberio, J., Irurozqui, E., Mendiburu, A., and Lozano, J. (2013). A distance-based ranking model estimation of distribution algorithm for the flowshop scheduling problem. *IEEE Trans. Evol. Comput.*, PP(99):1–1.

Chapman-Rothe, N., Curry, E., Zeller, C., Liber, D., Stronach, E., Gabra, H., Ghaem-Maghami, S., and Brown, R. (2012). Chromatin h3k27me3/h3k4me3 histone marks define gene sets in high-grade serous ovarian cancer that distinguish malignant, tumour-sustaining and chemo-resistant ovarian tumour cells. *Oncogene*.

Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R., and Adjaye, J. (2010). Computational analysis of genome-wide dna methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res.*, 20(10):1441–1450.

Chen, H. J., Edwards, R., Tucci, S., Bu, P., Milsom, J., Lee, S., Edelmann, W., Gümüs, Z. H., Shen, X., and Lipkin, S. (2012). Chemokine 25–induced signaling suppresses colon cancer invasion and metastasis. *J. Clin. Invest.*, 122(9):3184.

Chen, T., Lehre, P., Tang, K., and Yao, X. (2009). When is an estimation of distribution algorithm better than an evolutionary algorithm? In *Evolutionary Computation, 2009. CEC '09. IEEE Congress on*, pages 1470–1477.

Cheung, H., Lee, T., Davis, A., Taft, D., Rennert, O., and Chan, W. (2010). Genome-wide dna methylation profiling reveals novel epigenetically regulated genes and non-coding rnas in human testicular cancer. *Brit. J. Cancer*, 102(2):419–427.

Chin, L., Hahn, W. C., Getz, G., and Meyerson, M. (2011). Making sense of cancer genomic data. *Genes Dev.*, 25(6):534–555.

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H.-J., Yoon, D., Lee, M. H., Kim, D.-J., Park, M., et al. (2009). A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.*, 41:527–534.

Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. *Annu. Rev. Cell Dev.Biol.*, 26:721.

Cieply, B., Riley, P., Pifer, P. M., Widmeyer, J., Addison, J. B., Ivanov, A. V., Denvir, J., and Frisch, S. M. (2012). Suppression of the epithelial–mesenchymal transition by grainyhead-like-2. *Cancer Res.*, 72(9):2440–2453.

Clark, T., De Iorio, M., and Griffths, R. (2008). An evolutionary algorithm to find associations in dense genetic maps. *IEEE Trans. Evol. Comput.*, 12(3):297–306.

Cooper, L. R., Corne, D. W., and Crabbe, M. J. C. (2003). Use of a novel hill-climbing genetic algorithm in protein folding simulations. *Comput. Biol. Chem.*, 27(6):575–580.

Cordell, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, 10(6):392–404.

Das, D., Banerjee, N., and Zhang, M. Q. (2004). Interacting models of cooperative gene regulation. *Proc. Natl. Acad. Sci. U. S. A.*, 101(46):16234–16239.

Das, M. and Dai, H.-K. (2007). A survey of dna motif finding algorithms. *BMC Bioinformatics*, 8(Suppl 7):S21.

Deb, K. and Datta, R. (2010). A fast and accurate solution of constrained optimization problems using a hybrid bi-objective and penalty function approach. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–8.

Demircan, B., Dyer, L. M., Gerace, M., Lobenhofer, E. K., Robertson, K. D., and Brown, K. D. (2009). Comparative epigenomics of human and mouse mammary tumors. *Genes, Chromosomes and Cancer*, 48(1):83–97.

Ding, X.-Z., Kuszynski, C. A., El-Metwally, T. H., and Adrian, T. E. (1999). Lipoxygenase inhibition induced apoptosis, morphological changes, and carbonic anhydrase expression in human pancreatic cancer cells. *Biochem. Biophys. Res. Commun.*, 266(2):392–399.

Dohrmann, P. R., Butler, G., Tamai, K., Dorland, S., Greene, J. R., Thiele, D. J., and Stillman, D. J. (1992). Parallel pathways of gene regulation: homologous regulators swi5 and ace2 differentially control transcription of ho and chitinase. *Genes Dev.*, 6(1):93–104.

Dohrmann, P. R., Voth, W. P., and Stillman, D. J. (1996). Role of negative regulation in promoter specificity of the homologous transcriptional activators ace2p and swi5p. *Mol. Cell. Biol.*, 16(4):1746–1758.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–731.

Dumont, N., Wilson, M. B., Crawford, Y. G., Reynolds, P. A., Sigaroudinia, M., and Tlsty, T. D. (2008). Sustained induction of epithelial to mesenchymal transition activates dna methylation of genes silenced in basal-like breast cancers. *Proc. Natl. Acad. Sci. U. S. A.*, 105(39):14867–14872.

Dunham, I., Kundaje, A., Aldred, S., Collins, P., Davis, C., Doyle, F., Epstein, C., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Easwaran, H., Johnstone, S., Van Neste, L., Ohm, J., Mosbruger, T., Wang, Q., Aryee, M., Joyce, P., Ahuja, N., Weisenberger, D., Collisson, E., Zhu, J., Yegnasubramanian, S., Matsui, W., and Baylin, S. (2012). A dna hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Res.*, 22:837–849.

Ellis, L. M. and Hicklin, D. J. (2008). Vegf-targeted therapy: mechanisms of anti-tumour activity. *Nat. Rev. Cancer*, 8(8):579–591.

Ellsworth, R., Heckman, C., Seebach, J., Field, L., Love, B., Hooke, J., and Shriver, C. (2008). Identification of a gene expression breast cancer nodal metastasis profile. In *J. Clin. Oncol. (Meeting Abstracts)*, volume 26, page 1022.

Esteller, M. (2007). Epigenetic gene silencing in cancer: the dna hypermethylome. *Hum. Mol. Genet.*, 16(R1):R50–R59.

Esteller, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.*, 358(11):1148–1159.

Fadare, O. and Tavassoli, F. A. (2008). Clinical and pathologic aspects of basal-like breast cancers. *Nat. Clin. Pract. Oncol.*, 5(3):149–159.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G., Shen, R., Seshan, V., Mo, Q., Heguy, A., et al. (2011). Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci. Transl. Med.*, 3(75):75ra25.

Fang, J. Y. and Richardson, B. C. (2005). The mapk signalling pathways and colorectal cancer. *Lancet Oncol.*, 6(5):322–327.

Feero, W. G., Guttmacher, A. E., and Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, 363(2):166–176.

Fernandez, A., Garcia, S., Luengo, J., Bernado-Mansilla, E., and Herrera, F. (2010). Genetics-based machine learning for rule induction: State of the art, taxonomy, and comparative study. *IEEE Trans. Evol. Comput.*, 14(6):913–941.

Fiaschi, L., Garibaldi, J., and Krasnogora, N. (2010). A framework for the application of decision trees to the analysis of snps data. In *IEEE Symposium on*

*Computational Intelligence in Bioinformatics and Computational Biology 2009 (CIBCB 09)*, pages 106–113.

Fogel, G. B. and Corne, D. W. (2002). *Evolutionary computation in bioinformatics*. Morgan Kaufmann.

Fratkin, E., Naughton, B. T., Brutlag, D. L., and Batzoglou, S. (2006). Motifcut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics*, 22(14):e150–e157.

Grimm, D., Bauer, J., Pietsch, J., Infanger, M., Eucker, J., Eilles, C., and Schoenberger, J. (2011). Diagnostic and therapeutic use of membrane proteins in cancer cells. *Curr. Med. Chem.*, 18(2):176–190.

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

Handel, A. E., Ebers, G. C., and Ramagopalan, S. V. (2010). Epigenetics: molecular mechanisms and implications for disease. *Trends Mol. Med.*, 16(1):7–16.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664.

Heap, G. A., Trynka, G., Jansen, R. C., Bruinenberg, M., Swertz, M. A., Dinesen, L. C., Hunt, K. A., Wijmenga, C., Franke, L., et al. (2009). Complex nature of snp genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics*, 2(1):1.

Heidema, A., Feskens, E., Doevendans, P., Ruven, H., van Houwelingen, H., Mariman, E., and Boer, J. (2007). Analysis of multiple snps in genetic association studies: comparison of three multi-locus methods to prioritize and select snps. *Genet Epidemiol.*, 31(8):910–921.

Hill, V. K., Ricketts, C., Bieche, I., Vacher, S., Gentle, D., Lewis, C., Maher, E. R., and Latif, F. (2011). Genome-wide dna methylation profiling of cpg islands in breast cancer identifies novel genes associated with tumorigenicity. *Cancer Res.*, 71(8):2988–2999.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6(2):95–108.

Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80.

Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jönsson, G., Olsson, H., Borg, Å., and Ringnér, M. (2010). Molecular subtypes of breast cancer are associated with characteristic dna methylation patterns. *Breast Cancer Res.*, 12(3):R36.

Hong, C., Kim, Y. J., Moon, S., Shin, Y.-A., Cho, Y. S., and Lee, J.-Y. (2012). Karebrowser: Snp database of korea association resource project. *BMB Rep.*, 45(1):47–50.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R., Boks, M., van Eijk, K., van den Berg, L., and Ophoff, R. (2012). Aging effects on dna methylation modules in human brain and blood tissue. *Genome Biol.*, 13(10):R97.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of¡ i¿ cis¡/i¿-regulatory elements associated with groups of functionally related genes in¡ i¿ saccharomyces cerevisiae¡/i¿. *J. Mol. Biol.*, 296(5):1205–1214.

Huttenhower, C. and Hofmann, O. (2010). A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779.

Hvidsten, T. R., Wilczyński, B., Kryshtafovych, A., Tiuryn, J., Komorowski, J., and Fidelis, K. (2005). Discovering regulatory binding-site modules using rule-based learning. *Genome Res.*, 15(6):856–866.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nature Genet.*, 41(2):178–186.

Jackson, B. C., Carpenter, C., Nebert, D. W., Vasiliou, V., et al. (2010). Update of human and mouse forkhead box (fox) gene families. *Hum. Genomics*, 4:345–352.

Janssens, A. and van Duijn, C. (2008). Genome-based prediction of common diseases: advances and prospects. *Hum. Mol. Genet.*, 17(R2):R166–R173.

Jansson, E. Å., Are, A., Greicius, G., Kuo, I.-C., Kelly, D., Arulampalam, V., and Pettersson, S. (2005). The wnt/$\beta$-catenin signaling pathway targets ppar$\gamma$ activity in colon cancer cells. *Proc. Natl. Acad. Sci. U. S. A.*, 102(5):1460–1465.

Jeffery, I. B., Madden, S. F., McGettigan, P. A., Perriere, G., Culhane, A. C., and Higgins, D. G. (2007). Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, 23(3):298–305.

Jones, P. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492.

Jones, P. A. and Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4):683–962.

Jones, P. A. and Takai, D. (2001). The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070.

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.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489.

Kasturi, J. and Acharya, R. (2005). Clustering of diverse genomic data using information fusion. *Bioinformatics*, 21(4):423–429.

Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). Match: a tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res.*, 31(13):3576–3579.

Keleher, C. A., Passmore, S., and Johnson, A. (1989). Yeast repressor alpha 2 binds to its operator cooperatively with yeast protein mcm1. *Mol. Cell. Biol.*, 9(11):5228–5230.

Keles, S., van der Laan, M. J., and Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–1175.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. A., Niveleau, A., Cedar, H., et al. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.*, 38(2):149–153.

Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. F., Moran, J. L., et al. (2012). Exome sequencing and the genetic basis of complex traits. *Nat. Genet.*, 44(6):623–630.

Kim, J. H., Dhanasekaran, S. M., Prensner, J. R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M., et al. (2011). Deep sequencing reveals distinct patterns of dna methylation in prostate cancer. *Genome Res.*, 21(7):1028–1041.

Kim, S., Cho, H., Lee, D., and Webster, M. (2012). Association between snps and gene expression in multiple regions of the human brain. *Translational Psychiatry*, 2(5):e113.

Kitamura, T., Fujishita, T., Loetscher, P., Revesz, L., Hashida, H., Kizaka-Kondoh, S., Aoki, M., and Taketo, M. M. (2010). Inactivation of chemokine (cc motif) receptor 1 (ccr1) suppresses colon cancer liver metastasis by blocking accumulation of immature myeloid cells in a mouse model. *Proc. Natl. Acad. Sci. U. S. A.*, 107(29):13063–13068.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70.

Kouskoumvekaki, I., Shublaq, N., and Brunak, S. (2013). Facilitating the use of large-scale biological data and tools in the era of translational bioinformatics. *Brief. Bioinform.*, page bbt055.

Krebs, A., Frontini, M., and Tora, L. (2008). Gpat: retrieval of genomic annotation from large genomic position datasets. *BMC Bioinformatics*, 9(1):533.

Kumar, M., Husian, M., Upreti, N., and Gupta, D. (2010). Genetic algorithm: Review and application. *Int. J. Inform. Tech. Knowl. Manag.*, 2(2):451–454.

kumar Singh, A., Singh, R., Naz, F., Chauhan, S. S., Dinda, A., Shukla, A. A., Gill, K., Kapoor, V., and Dey, S. (2012). Structure based design and synthesis of peptide inhibitor of human lox-12: in vitro and in vivo analysis of a novel therapeutic agent for breast cancer. *PloS One*, 7(2):e32521.

Lacroix, M. (2006). Significance, detection and markers of disseminated breast cancer cells. *Endocr.-Relat. Cancer*, 13(4):1033–1067.

Laird, P. (2010). Principles and challenges of genomewide dna methylation analysis. *Nat. Rev. Genet.*, 11(3):191–203.

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, 274(5287):536–539.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Li, X., Wu, L., Corsa, C. A. S., Kunkel, S., and Dou, Y. (2009). Two mammalian mof complexes regulate transcription activation by distinct mechanisms. *Mol. Cell*, 36(2):290–301.

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

Liu, F., Tsai, J.-P., Chen, R., Chen, S., and Shih, S. H. (2004). Fmga: finding motifs by genetic algorithm. In *Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on*, pages 459–466.

Lombaerts, M., Van Wezel, T., Philippo, K., Dierssen, J., Zimmerman, R., Oosting, J., van Eijk, R., Eilers, P., van De Water, B., Cornelisse, C., et al. (2006). E-cadherin transcriptional downregulation by promoter methylation but not mutation is related to epithelial-to-mesenchymal transition in breast cancer cell lines. *Brit. J. Cancer*, 94(5):661–671.

Lydall, D., Ammerer, G., and Nasmyth, K. (1991). A new role for mcm1 in yeast: cell cycle regulation of sw15 transcription. *Genes Dev.*, 5(12b):2405–2419.

MacKay, V. L., Mai, B., Waters, L., and Breeden, L. L. (2001). Early cell cycle box-mediated transcription ofcln3 and swi4 contributes to the proper timing of the g1-to-s transition in budding yeast. *Mol. Cell. Biol.*, 21(13):4140–4148.

Mahony, S., Hendrix, D., Golden, A., Smith, T. J., and Rokhsar, D. S. (2005). Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–1814.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., et al. (2003). Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34(suppl 1):D108–D110.

Mayor, R., Muñoz, M., Coolen, M. W., Custodio, J., Esteller, M., Clark, S. J., and Peinado, M. A. (2011). Dynamics of bivalent chromatin domains upon drug induced reactivation and resilencing in cancer cells. *Epigenetics*, 6(9):1138–1148.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, 9(5):356–369.

McGarvey, K. M., Van Neste, L., Cope, L., Ohm, J. E., Herman, J. G., Van Criekinge, W., Schuebel, K. E., and Baylin, S. B. (2008). Defining a chromatin pattern that characterizes dna-hypermethylated genes in colon cancer cells. *Cancer Res.*, 68(14):5753–5759.

Mead, J., Zhong, H., Acton, T. B., and Vershon, A. K. (1996). The yeast alpha2 and mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. *Mol. Cell. Biol.*, 16(5):2135–2143.

Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., Baylin, S. B., and Sidransky, D. (1995). 52 cpg island methylation is associated with transcriptional silencing of the tumour suppressor p16/cdkn2/mts1 in human cancers. *Nat. Med.*, 1(7):686–692.

Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nat. Rev. Genet.*, 11(1):31–46.

Moore, J., Asselbergs, F., and Williams, S. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455.

Moore, J. and White, B. (2007). *Genome-wide association studies for common dis-*

*eases and complex traits.* Genetic Programming Theory and Practice IV. Springer, New York.

Morillon, A., O'Sullivan, J., Azad, A., Proudfoot, N., and Mellor, J. (2003). Regulation of elongating rna polymerase ii by forkhead transcription factors in yeast. *Science*, 300(5618):492–495.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662.

Morris, M., Gentle, D., Abdulrahman, M., Clarke, N., Brown, M., Kishida, T., Yao, M., Teh, B., Latif, F., and Maher, E. R. (2008). Functional epigenomics approach to identify methylated candidate tumour suppressor genes in renal cell carcinoma. *Br. J. Cancer*, 98(2):496–501.

Morrow, B. E., Johnson, S. P., and Warner, J. R. (1989). Proteins that bind to the yeast rdna enhancer. *J. Biol. Chem.*, 264(15):9061–9068.

Murrell, A., Rakyan, V. K., and Beck, S. (2005). From genome to epigenome. *Hum. Mol. Genet.*, 14(suppl 1):R3–R10.

Namkung, J., Nam, J., and Park, T. (2007). Identification of expression quantitative trait loci by the interaction analysis using genetic algorithm. *BMC Proc.*, 1(1):S69.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527.

Nguyen, H. D., Yoshihara, I., Yamamori, K., and Yasunaga, M. (2002). A parallel hybrid genetic algorithm for multiple protein sequence alignment. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, pages 309–314. IEEE.

Notredame, C. and Higgins, D. G. (1996). Saga: sequence alignment by genetic algorithm. *Epigenetics*, 24(8):1515–1524.

Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K., and Wegener, I. (2007). Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 23(24):3280–3288.

Ohgami, R. S., Ma, L., Ren, L., Weinberg, O. K., Seetharam, M., Gotlib, J. R., and Arber, D. A. (2012). Dna methylation analysis of alox12 and gstm1 in acute myeloid leukaemia identifies prognostically significant groups. *Brit. J. Haematol.*, 159(2):182–190.

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., and Gilad, Y. (2011). A genome-wide study of dna methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.*, 7(2):e1001316.

Pal, S., Bandyopadhyay, S., and Ray, S. (2006). Evolutionary computation in bioinformatics: a review. *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, 36(5):601–615.

Palsson, B. and Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.*, 6(11):787–789.

Park, P. J., Butte, A. J., and Kohane, I. S. (2002). Comparing expression profiles of genes with similar promoter regions. *Bioinformatics*, 18(12):1576–1584.

Parsons, D. W., Wang, T.-L., Samuels, Y., Bardelli, A., Cummins, J. M., DeLong, L., Silliman, N., Ptak, J., Szabo, S., Willson, J. K., et al. (2005). Colorectal cancer: mutations in a signalling pathway. *Nature*, 436(7052):792–792.

Pelikan, M. (2006). Implementation of the dependency-tree estimation of distribution algorithm in c++.

Pelikan, M., Tsutsui, S., and Kalapala, R. (2007). Dependency trees, permutations, and quadratic assignment problem. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, GECCO '07, pages 629–629, New York, NY, USA. ACM.

Pidgeon, G. P., Kandouz, M., Meram, A., and Honn, K. V. (2002). Mechanisms controlling cell cycle arrest and induction of apoptosis after 12-lipoxygenase inhibition in prostate cancer cells. *Cancer Res.*, 62(9):2721–2727.

Pidgeon, G. P., Tang, K., Cai, Y. L., Piasentin, E., and Honn, K. V. (2003). Overexpression of platelet-type 12-lipoxygenase promotes tumor cell survival by enhancing $\alpha$v$\beta$3 and $\alpha$v$\beta$5 integrin expression. *Cancer Res.*, 63(14):4258–4267.

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29(2):153–159.

Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. *Nat. Biotech.*, 28(10):1057–1068.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.*, 12(5):R68.

Rao, X., Evans, J., Chae, H., Pilrose, J., Kim, S., Yan, P., Huang, R., Lai, H., Lin, H., Liu, Y., et al. (2013). Cpg island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, 32:4519—-4528.

Rauch, T. A. and Pfeifer, G. P. (2010). Dna methylation profiling using the methylated-cpg island recovery assay (mira). *Methods*, 52(3):213–217.

Reddington, J. P., Perricone, S. M., Nestor, C. E., Reichmann, J., Youngson, N. A., Suzuki, M., Reinhardt, D., Dunican, D. S., Prendegast, J. G., Mjoseng, H., et al. (2013). Redistribution of h3k27me3 upon dna hypomethylation results in derepression of polycomb-target genes. *Genome Biol.*, 14(3):R25.

Rhee, J.-K., Joung, J.-G., Chang, J.-H., Fei, Z., and Zhang, B.-T. (2009). Identification of cell cycle-related regulatory motifs using a kernel canonical correlation analysis. *BMC Genomics*, 10(Suppl 3):S29.

Rhee, J.-K., Kim, K., Chae, H., Evans, J., Yan, P., Zhang, B.-T., Gray, J., Spellman, P., Huang, T. H.-M., Nephew, K. P., et al. (2013). Integrated analysis of genome-wide dna methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Res.*, 41(18):8464–8474.

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T. R., Ghosh, D., and Chinnaiyan, A. M. (2005). Mining for regulatory programs in the cancer transcriptome. *Nat. Genet.*, 37(6):579–583.

Ring, A., Smith, I. E., and Dowsett, M. (2004). Circulating tumour cells in breast cancer. *Lancet Oncol.*, 5(2):79–88.

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F., and Moore, J. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, 69(1):138–147.

Robertson, K. D. (2005). Dna methylation and human disease. *Nat. Rev. Genet.*, 6(8):597–610.

Robinson, M. D., Stirzaker, C., Statham, A. L., Coolen, M. W., Song, J. Z., Nair, S. S., Strbenac, D., Speed, T. P., and Clark, S. J. (2010). Evaluation of affinity-based genome-wide dna methylation data: effects of cpg density, amplification bias, and copy number variation. *Genome Res.*, 20(12):1719–1729.

Rodriguez, J., Muñoz, M., Vives, L., Frangou, C. G., Groudine, M., and Peinado, M. A. (2008). Bivalent domains enforce transcriptional memory of dna methylated genes in cancer cells. *Proc. Natl. Acad. Sci. U. S. A.*, 105(50):19809–19814.

Roskoski Jr, R. (2004). The erbb/her receptor protein-tyrosine kinases and cancer. *Biochem. Biophys. Res. Commun.*, 319(1):1–11.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2004). Exploring interactions in high dimensional genomic data: An overview of logic regression, with applications. *J. Multivar. Anal.*, 90:178–195.

Ruike, Y., Imanaka, Y., Sato, F., Shimizu, K., and Tsujimoto, G. (2010). Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-dna immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, 11(1):137.

Sadikovic, B., Al-Romaih, K., Squire, J., and Zielenska, M. (2008). Cause and consequences of genetic and epigenetic alterations in human cancer. *Curr. Genomics*, 9(6):394–408.

Sakai, T., Toguchida, J., Ohtani, N., Yandell, D. W., Rapaport, J. M., and Dryja, T. P. (1991). Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am. J. Hum. Genet.*, 48(5):880.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32(suppl 1):D91–D94.

Sandoval, J. and Esteller, M. (2012). Cancer epigenomics: beyond genomics. *Curr. Opin. Genet. Dev.*, 22(1):50–55.

Santana, R., Mendiburu, A., Zaitlen, N., Eskin, E., and Lozano, J. (2010). Multi-marker tagging single nucleotide polymorphism selection using estimation of distribution algorithms. *Artif. Intell. Med.*, 50(3):193–201.

Savage, K., Lambros, M. B., Robertson, D., Jones, R. L., Jones, C., Mackay, A., James, M., Hornick, J. L., Pereira, E. M., Milanezi, F., et al. (2007). Caveolin 1 is overexpressed and amplified in a subset of basal-like and metaplastic breast carcinomas: a morphologic, ultrastructural, immunohistochemical, and in situ hybridization analysis. *Clin. Cancer Res.*, 13(1):90–101.

Segal, E., Yelensky, R., and Koller, D. (2003). Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19(suppl 1):i273–i282.

Segditsas, S. and Tomlinson, I. (2006). Colorectal cancer and genetic alterations in the wnt pathway. *Oncogene*, 25(57):7531–7537.

Sharma, S., Kelly, T. K., and Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36.

Shelke, K., Jayaraman, S., Ghosh, S., and Valadi, J. (2013). Hybrid feature selection and peptide binding affinity prediction using an eda based algorithm. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 2384–2389.

Shim, V., Tan, K., Chia, J., and Al Mamun, A. (2013). Multi-objective optimization with estimation of distribution algorithm in a noisy environment. *Evol. Comput.*, 21(1):149–177.

Sigaud, O. and Wilson, S. (2007). Learning classifier systems: a survey. *Soft Comput.*, 11(11):1065–1078.

Simmer, F., Brinkman, A., Assenov, Y., Matarese, F., Kaan, A., Sabatino, L., Villanueva, A., Huertas, D., Esteller, M., Lengauer, T., Bock, C., Colantuoni, V., Altucci, L., and Stunnenberg, H. (2012). Comparative genome-wide dna methylation analysis of colorectal tumor and matched normal tissues. *Epigenetics*, 7(12):1355–1367.

Simon, D. (2013). *Evolutionary Optimization Algorithms: Biologically-Inspired and Population-Based Approaches to Computer Intelligence.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., et al. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708.

Singh, A., Kant, S., Parshad, R., Banerjee, N., and Dey, S. (2011). Evaluation of human lox-12 as a serum marker for breast cancer. *Biochem. Biophys. Res. Commun.*, 414(2):304–308.

Slattery, M. L., Lundgreen, A., and Wolff, R. K. (2012). Map kinase genes and colon and rectal cancer. *Carcinogenesis*, 33(12):2398–2408.

Sloan, E. K., Stanley, K. L., and Anderson, R. L. (2004). Caveolin-1 inhibits breast cancer growth and metastasis. *Oncogene*, 23(47):7893–7897.

Spano, J., Fagard, R., Soria, J.-C., Rixe, O., Khayat, D., and Milano, G. (2005). Epidermal growth factor receptor signaling in colorectal cancer: preclinical data and therapeutic perspectives. *Ann. Oncol.*, 16(2):189–194.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297.

Sproul, D., Nestor, C., Culley, J., Dickson, J. H., Dixon, J. M., Harrison, D. J., Meehan, R. R., Sims, A. H., and Ramsahoye, B. H. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 108(11):4364–4369.

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383.

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

Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R., Carr, J. M., Khrebtukova, I., Luo, S., Zhang, L., et al. (2011). Integrated analysis of gene expression, cpg island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, 6(2):e17490.

Suzuki, M. M. and Bird, A. (2008). Dna methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, 9(6):465–476.

Szymczak, S., Biernacka, J., Cordell, H., Gonzalez-Recio, O., Konig, I., Zhang, H., and Sun, Y. (2009). Machine learning in genome-wide association studies. *Genet. Epidemiol.*, 33(Suppl 1):S51–57.

Takai, D. and Jones, P. A. (2002). Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.*, 99(6):3740–3745.

Tan, A. C., Jimeno, A., Lin, S. H., Wheelhouse, J., Chan, F., Solomon, A., Rajeshkumar, N., Rubio-Viqueira, B., and Hidalgo, M. (2009). Characterizing dna methylation patterns in pancreatic cancer genome. *Mol. Oncol.*, 3(5):425–438.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.*, 22(3):281–285.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E., and Schork, N. (2011). The importance of phase information for human genomics. *Nat. Rev. Genet.*, 12(3):213–223.

Thiery, J. P. (2002). Epithelial–mesenchymal transitions in tumour progression. *Nat. Rev. Cancer*, 2(6):442–454.

Thiery, J. P., Acloque, H., Huang, R. Y., and Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *Cell*, 139(5):871–890.

Toft, D. J. and Cryns, V. L. (2011). Minireview: basal-like breast cancer: from molecular profiles to targeted therapies. *Mol. Endocrinol.*, 25(2):199–211.

Toperoff, G., Aran, D., Kark, J. D., Rosenberg, M., Dubnikov, T., Nissan, B., Wainstein, J., Friedlander, Y., Levy-Lahad, E., Glaser, B., et al. (2012). Genome-wide survey reveals predisposing diabetes type 2-related dna methylation variations in human peripheral blood. *Hum. Mol. Genet.*, 21(2):371–383.

Tsai, H.-K., Lu, H. H.-S., and Li, W.-H. (2005). Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, 102(38):13532–13537.

Ueki, T., Walter, K. M., Skinner, H., Jaffee, E., Hruban, R. H., and Goggins, M. (2002). Aberrant cpg island methylation in cancer cell lines arises in the primary cancers from which they were derived. *Oncogene*, 21(13):2114–2117.

Unger, R. and Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, 231(1):75–81.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Vershon, A. K. and Johnson, A. D. (1993). A short, disordered protein region mediates interactions between the homeodomain of the yeast $\alpha 2$ protein and the mcm1 protein. *Cell*, 72(1):105–112.

Walker, B. A., Wardell, C. P., Chiecchio, L., Smith, E. M., Boyd, K. D., Neri, A., Davies, F. E., Ross, F. M., and Morgan, G. J. (2011). Aberrant global methylation patterns affect the molecular pathogenesis and prognosis of multiple myeloma. *Blood*, 117(2):553–562.

Wang, B., Chen, P., Zhang, J., Zhao, G., and Zhang, X. (2010). Inferring protein-protein interactions using a hybrid genetic algorithm/support vector machine method. *Protein Pept. Lett.*, 17(9):1079–1084.

Wang, R., Purshouse, R. C., and Fleming, P. J. (2013). On finding well-spread pareto optimal solutions by preference-inspired co-evolutionary algorithm. In *Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference*, GECCO '13, pages 695–702, New York, NY, USA. ACM.

Werth, M., Walentin, K., Aue, A., Schönheit, J., Wuebken, A., Pode-Shakked, N., Vilianovitch, L., Erdmann, B., Dekel, B., Bader, M., et al. (2010). The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex. *Development*, 137(22):3835–3845.

Winder, T. and Lenz, H.-J. (2010). Vascular endothelial growth factor and epidermal growth factor signaling pathways as therapeutic targets for colorectal cancer. *Gastroenterology*, 138(6):2163–2176.

Xiang, X., Deng, Z., Zhuang, X., Ju, S., Mu, J., Jiang, H., Zhang, L., Yan, J., Miller, D., and Zhang, H.-G. (2012). Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. *PloS One*, 7(12):e50781.

Xie, X., Lu, J., Kulbokas, E., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and $3'$ utrs by comparison of several mammals. *Nature*, 434(7031):338–345.

Yamanishi, Y., Vert, J.-P., Nakaya, A., and Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19(suppl 1):i323–i330.

Yang, P., Ho, J., Zomaya, A., and Zhou, B. (2010). A genetic ensemble approach for gene-gene interaction identification. *BMC Bioinformatics*, 11:524.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.*, 39(17):7415–7427.

Yuan, T. and Cantley, L. (2008). Pi3k pathway alterations in cancer: variations on a theme. *Oncogene*, 27(41):5497–5510.

Zhang, J., Zhan, Z.-h., Lin, Y., Chen, N., Gong, Y.-j., Zhong, J.-h., Chung, H. S., Li, Y., and Shi, Y.-h. (2011). Evolutionary computation meets machine learning: A survey. *IEEE Comput. Intell. Mag.*, 6(4):68–75.

Zhang, S., Liu, X., Zhang, Y., Cheng, Y., and Li, Y. (2013). Rnai screening identifies kat8 as a key molecule important for cancer cell survival. *Int. J. Clin. Exp. Pathology*, 6(5):870.

Zhong, H., McCord, R., and Vershon, A. K. (1999). Identification of target sites of the $\alpha$2–mcm1 repressor complex in the yeast genome. *Genome Res.*, 9(11):1040–1047.

Zhou, A., Zhang, Q., and Jin, Y. (2009). Approximating the set of pareto-optimal solutions in both the decision and objective spaces by an estimation of distribution algorithm. *IEEE Trans. Evol. Comput.*, 13(5):1167–1189.

Zhuang, J., Jones, A., Lee, S.-H., Ng, E., Fiegl, H., Zikan, M., Cibula, D., Sargent, A., Salvesen, H. B., Jacobs, I. J., et al. (2012). The dynamics and prognostic potential of dna methylation changes at stem cell gene loci in women's cancer. *PloS Genet.*, 8(2):e1002517.

Zuo, T., Tycko, B., Liu, T.-M., Lin, H.-J. L., and Huang, T. H. (2009). Methods in dna methylation profiling. *Epigenomics*, 1(2):331–345.

# 초 록

생명과학 연구의 기본적 목표 중 하나는 생물학적 인자들과 표현형의 복잡한 관계를 이해하고, 표현형에 영향을 미치는 다양한 인자들을 밝히는 것이다. 특히 유전체 서열은 유전자 발현이나 질병 민감도 등의 표현형을 결정하는 데에 있어서 중요한 역할을 한다. 따라서 유전체 서열 기반 정보에 대한 연구는 생물학적 기작을 이해하기 위해 필수적이다. 기존의 유전체 서열 관련 연구는 주로 생체 내 기작에 중요한 영향을 미치는 하나의 인자를 찾는 것에 집중되어 있었다. 최근 대용량 생물학 데이터 생산 기술의 발전으로 인해 전역 유전체 수준에서 유전적 변이를 분석하고 질병의 원인을 찾고자 하는 시도가 가능하게 되었지만, 거대한 탐색 공간과 계산 복잡도로 인해 여전히 다중 인자들의 고차 관계를 탐색하여 분석하는 것은 쉬운 일이 아니다.

본 논문에서는 진화 연산과 통계적 학습 방법을 결합하여 다중 인자 상호 작용을 탐색할 수 있는 효과적인 방법들을 제안한다. 본 논문의 방법들은 다양한 전역 유전체 서열 분석 문제에서 상호 연관된 인자 조합과 기능적 모듈의 탐색을 목적으로 한다. 우선 통계적 학습 방법을 이용하여 유전자 발현 조절에 함께 영향을 주는 서열 조각 및 DNA 메틸화 영역을 탐색한다. 이후 인간 유전체와 같이 많은 수의 인자들을 가진 고차원의 서열 데이터 분석을 위해 진화 연산 개념을 도입한다. 본 논문에서 사용된 방법은 학습 데이터를 이용한 기계 학습 기술을 기반으로 하여 진화 연산 과정에서 문제 공간을 효과적으로 탐색한다. 이를 통해 계산학적으로 복잡한 최적화 문제에서 답이 될 수 있는 후보군들을 찾아가는 것이 가능하다. 유전체 및 후성유전체 데이터를 이용한 실험 결과는 본 논문에서 사용된 진화 연산 기반 방법이 질병과 연관된 고차 상호 관계를 발견할 수 있다는 것을 보인다. 따라서 본 논문의 연구는 유전체 및 후성유전체 연구에서 서열 기반 인자들 간의 복잡한 상호작용을 분석할 수 있는 유용한 방법이 될 수 있을 것이다.

**Keywords:** **고차상호작용, 진화연산, 유전체 서열 분석, 기계학습, 유전체학, 후성유전체학**

**학번: 2004-20623**