의학박사 학위논문

Graph- and kernel-based integrative analyses of multi-layers of heterogeneous genomic data

이종 다계층 유전체 정보의 그래프 기반 통합 및 커널 기반 통합 분석 연구

2013 년 2 월

서울대학교 대학원
의학과 분자유전체의학과정
김 도 균

A thesis of the Degree of Doctor of Philosophy

이종 다계층 유전체 정보의 그래프 기반 통합 및 커널 기반 통합 분석 연구

Graph- and kernel-based integrative analyses of multi-layers of heterogeneous genomic data

February 2013

The Department of Molecular and Genomic Medicine,
Seoul National University
College of Medicine
Do Kyoon Kim

Graph- and kernel-based integrative analyses of multi-layers of heterogeneous genomic data

by
Do Kyoon Kim

A thesis submitted to the Department of Molecular and Genomic Medicine in partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Bioinformatics at Seoul National University College of Medicine

February 2013

Approved by Thesis Committee:

Professor 박 웅 양 Chairman
Professor 김 주 한 Vice chairman
Professor 문 병 로
Professor 전 주 홍
Professor 고 인 송

ABSTRACT

Introduction: Cancer is a complex disease, which can be dysregulated through multiple mechanisms. Therefore, no single level of genomic data fully elucidates tumor behavior since there are many genomic variations within or between levels in a biological system such as copy number variants, DNA methylation, alternative splicing, miRNA regulation, post translational modification, *etc*. Nowadays, a number of heterogeneous types of data have become more available (i.e., TCGA, the Cancer Genome Atlas) which are generated from multiple molecular levels of omics dimensions from genome to phenome.

Methods: Given multi-levels of data, information from a level to another may lead to some clues that help to uncover an unknown biological knowledge. Thus, integration of different levels of data can aid in extracting new knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of genomic data. In the meantime, it is expected that the next attempt is more focused on how to utilize the information from inter-relation, the relation between different levels: from the genome level to epigenome, transcriptome, proteome, and further stretched to the phenome level. In this study, the prototypes of the research schemes for integrative analysis of multi-layers and heterogeneous genomic data were introduced and discussed.

Results: These schemes were exemplified based on the pilot experimental results on the prediction problem of cancer clinical outcomes using the TCGA

data. For glioblastoma multiforme, all clinical outcomes had a better the area under the curve (AUC) of receiver operating characteristic when integrating multi-layers of genomic data, 0.876 for survival to 0.832 for recurrence. Moreover, the better AUCs were achieved from the integration approach for all clinical outcomes in ovarian cancer as well, ranging from 0.787 to 0.893. In addition, based on our results, an accuracy of prediction model with inter-relationship increases because of incorporation of information fused over genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression.

Conclusions: I found that the opportunity for success in prediction of clinical outcomes in cancer was increased when the prediction was based on the integration of multi-layers of genomic data and genomic knowledge. This study is expecting to improve comprehension of the molecular pathogenesis and underlying biology of both cancer types.

Keywords: Integrative analysis, Multi-layers of genomic data, Clinical outcome prediction, Glioblastoma multiforme, Serous cystadenocarcinoma

Student number: 2006-22113

CONTENTS

LIST OF TABLES AND FIGURES

General Introduction

Chapter 1

Chapter 2

Chapter 3

Chapter 4

GENERAL INTRODUCTION

Understanding of the molecular basis of cancer brings many benefits for predicting clinical outcomes of cancer and for determining the corresponding best treatment. Since cancer is related to alterations in the genes that control normal cell growth and death, molecular-based diagnostics are promising in that they may provide more opportunities for objective, precise, and systematic predictions on cancer. Data at the multiple molecular levels, generated from all levels of omic' dimensions from genome to phenome (Fig. 1), have recently become more available. At the genome level, copy number variants have attracted considerable attentions, since alterations of genomic DNA can be explored by expanding the scope of view to a larger region of the genome or to chromosomes. At the epigenome level, data from DNA methylation, which plays a crucial role in the control of gene activity, is of interest, while at the level of the transcriptome, gene expression and microRNA (miRNA) are the most representative datasets. DNA microarrays have already been widely used for the classification of tumor subtypes or clinical outcomes for the diagnosis, treatment, or prognosis of cancer for many years (1-6). More recently, miRNA has become available for understanding the inhibition of expression on target mRNAs in gene regulatory networks.

There have been attempts at cancer classification based on a set of miRNA, copy number alterations (CNA), and DNA methylation (7-11). Despite these efforts, however, it still remains difficult to elucidate the cancer phenotypes because the cancer genome is neither simple nor independent but rather

complicated and dysregulated by multiple molecular mechanisms (12-13). For example, the cancer genome is related to mutations in coding and non-coding sequences, changes in the DNA structure and copy number, DNA methylation and histone modification, and miRNA regulation. Those possibilities lead to many alternative forms of cause-and-result in transcription, translation, post-translational modification, and eventually, gene and protein functions (Fig. 1). Thus, no single level of genomic data will be sufficient to comprise all the information in the mechanism, and hence, a consideration of the layered processes in biological systems through incorporation of multiple levels of genomic data might provide much more reasonable prediction of cancer phenotypes.

The Cancer Genome Atlas (TCGA) is a collaborative initiative to improve understanding of cancer using existing large-scale whole-genome technologies. The TCGA research network lately published many notable papers on glioblastoma patients concerning an interim analysis of DNA sequencing, copy number, gene expression, and DNA methylation data (14-18), and discovery of links between cancer subtypes and different neural lineages with gene expression (19). While the TCGA opens many opportunities to researchers to deepen the knowledge of the molecular basis of cancer (19-26), it is particularly important to access multiple data sources as I propose here.

In this research, I propose an integrated framework that uses multi-level genomic data sources for the molecular-based classification of clinical outcomes in cancer. Understanding the molecular pathogeneses and

underlying biology for several cancers is expected to provide guidance for improved prognostic indicators and effective therapies.

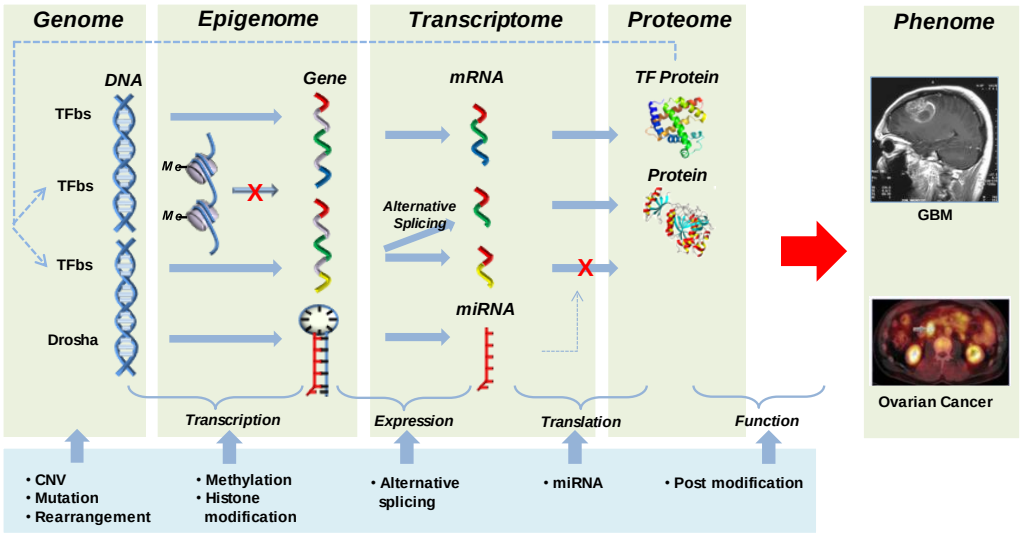


Figure 1. Multi-layers of genomic data in a biological system from genome, epigenome, transcriptome and proteome to phenome. There are many genomic variations within or between levels: Copy number variant (CNV), sequence mutation, and genomic rearrangement in genome level; DNA methylation and histone modification in epigenome level; alternative splicing and miRNA regulation in transcriptome level; post translational modification in proteome level. TF, transcription factor; TFbs, transcription factor binding site; Me, methylation; Drosha, a class 2 RNase III enzyme responsible for initiating the processing of microRNA.

In computational biology, this work will be a pioneering attempt to predict the cancer phenotype based on the underlying complex biological

mechanisms. From individual TCGA data sources, empirical comparisons were conducted at each level of genomic data; to deduce possible biological implications based on the results of the relative contribution of each piece of data to increase prediction accuracy. In addition, I assume that accuracy of prediction increases because of incorporation of information fused over heterogeneous biological data sources, providing an enhanced global view on cancer mechanisms in an intermediate integration manner (Fig. 2). The approach of intermediate integration has the advantage that a model is trained by weighing both multi-layers of genomic data simultaneously based on a kernel or graph levels. Instead of two independent hypotheses that have to be combined afterwards, these results into one prediction for each patient and only one hypothesis have to be formed. Moreover, when integrating multi-layers of heterogeneous genomic data, it is desirable that a framework is capable of containing the inter-relationships between sample features belonging to different layers of the biological system. Finally, in order to explain the phenotype of complex diseases, the integrative framework should be extended to incorporate with much genomic knowledge such as pathway, Gene Ontology, chromosomal positional geneset, motif geneset, protein-protein interaction, etc.

Several approaches to multiple data integration have been applied to protein function prediction such as the kernel-based integration framework (27-29), the Relevance Vector Machine (RVM) approach (30), and a Bayesian model (31). In recent years, Shin *et al*. developed an integration method of protein networks based on graph-based semi supervised learning (SSL), which is

halfway between supervised and unsupervised learning (32-33). Although, any of the above mentioned methods could be used to implement the proposed idea, the latter one is employed in this study, taking advantage of computational efficiency and representational ease for the biological system. The learning time of graph-based SSL is nearly linear with the number of graph edges, which in most biological networks is few, while the accuracy remains comparable to the kernel-based methods that suffer from the relative disadvantage of a longer learning time (33-34). In addition, the interpretation of biological phenomena can be improved because of the graph data structure (35-37), which naturally fits into the graph based SSL.

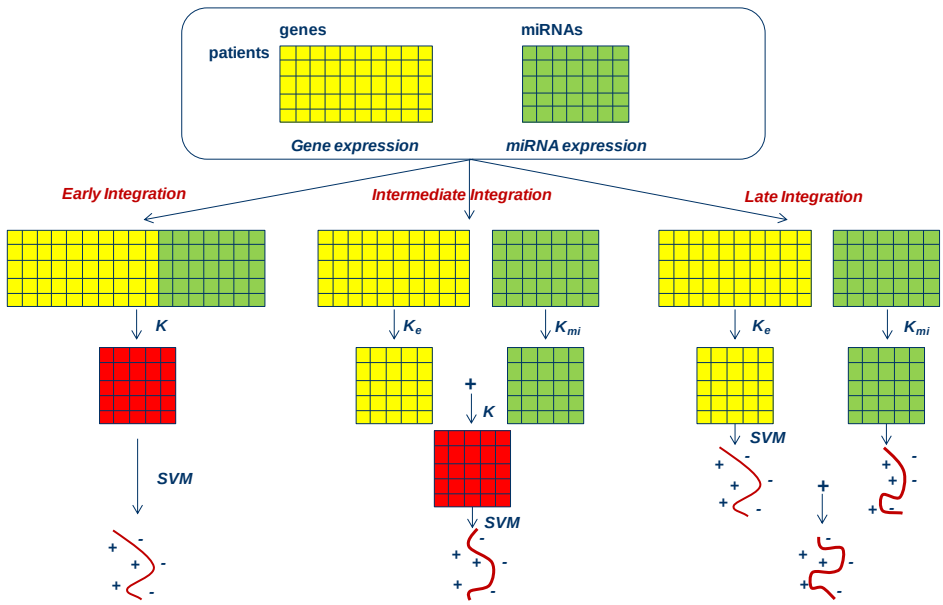


Figure 2. Data integration in machine learning.

This study can be categorized into four chapters of sub-studies.

1. Genomic data comparison: Which data is more informative?

Recently, various types of genomic data from cancer patients have become available thanks to the collaborative initiatives in better understanding of cancer. With abundance in genomic/clinical data, the question that bioinformaticians often encounter is which data is more informative. To wet-lab analysts, it concerns data generation that requires highly cost/time-demanding work and experienced facilities. To dry-lab analysts, it concerns selection of appropriate data source for more accurate prediction, avoiding unnecessary waste of computational resource. To provide a preliminary insight on the question, this study compares different types of genomic data using the state-of-the-art machine learning algorithm, Semi-Supervised Learning.

2. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction

There have been many attempts in cancer clinical outcome prediction by suing a dataset from a number of molecular layers of biological system. Despite these efforts, however, it still remains difficult to elucidate the cancer phenotypes because the cancer genome is neither simple nor independent but rather complicated and dysregulated by multiple molecular mechanisms. Recently, heterogeneous types of genomic data, generated from all molecular levels of 'omic' dimensions from genome to phenome, for instance, *copy number variants* at the genome level, *DNA methylation* at the epigenome level,

and *gene expression* and *microRNA* at the transcriptome level, have become available. In this study, I propose an integrated framework that uses multi-layers of heterogeneous genomic data for prediction of clinical outcomes in brain cancer (Glioblastoma multiforme, GBM) and ovarian cancer (Serous cystadenocarcinoma, OV).

3. Combining multi-layers of genomic data and inter-relationship between different layers of genomic features

The limitation of previous study is integration with multi-layers of genomic data for cancer clinical outcome prediction without considering of inter-relationship information between them. There are possible relationships between the sample features (attributes) belonging to different layers of genomic data such as 'miRNA-target genes,' 'copy number alteration region-genes located in the alteration region,' 'DNA methylation site-specific genes regulated by promoter regions,' etc. Therefore, when integrating multiple genomic data, it will be desirable that a framework will be capable of containing the inter-relationships between sample features belonging to different layers of the biological system. This study can be categorized into three types of sub-studies.

4. Knowledge bootstrapping: a graph-based integration with multi-omics data and genomic knowledge

Finally, in order to explain the phenotype of complex diseases, it is better way to incorporate the genomic knowledge when integrating multi-layers and heterogeneous genomic data. Several methods with integrating genomic

knowledge such as pathways or protein-protein interaction networks based on gene expression data have been developed to overcome variability of diagnostic or prognostic predictors and to increase their performances However, none of previous studies provided the integrative framework for multi-omics data and genomic knowledge. Here, I propose a new integrative framework for multi-omics and genomic knowledge in order to better explain the phenotype of complex diseases.

CHAPTER 1

Genomic data comparison: Which data is more informative?

INTRODUCTION

With abundance in genomic and clinical data, now the question that bioinformaticians often encounter is cast on which data is more informative. To wet-lab analysts, it concerns data generation that requires highly cost/time-demanding work and experienced facilities. To dry-lab analysts, it concerns selection of appropriate data source for accurate prediction, avoiding unnecessary waste of computational resource. For instance, the gene expression data analysis often incurs intractable computational complexity due to its high dimensionality.

To provide a preliminary insight on the question, this study compares different types of genomic data based on the classification problem of initial vs. recurrent tumor in GBM. GBM is the most common and aggressive primary brain tumor in adults (38), and notorious for its tendency to recur (39). Despite recent advances in the molecular pathology of GBM, the underling molecular mechanisms associated with clinical outcome are still poorly understood (38, 40).

MATERIALS AND METHODS

Data

Table 1-1 shows the four types of GBM related genomic data, CNA, methylation, miRNA, and gene expression (TCGA data portal, http://tcga-data.nci.nih.gov/). The total 159 patients' records were available across the four data sets (N=159), in which 39 were recurred tumor (y=-1) while the remaining were initial tumor (y=1). Technically, the data-setup of our experiment for binary classification can be rephrased as $\{x_n, y_n\}_{n=1}^{N}$ where $x_n \in R^d$ (d is the number of features and N is the number of records) and $y_n \in \{-1,1\}$. The initial tumor occurrence was defined referring to the values of procedure type and pretreatment history, "surgical resection" and "no pre-history," respectively. On the other hand, the recurred tumor was defined, according to the phenotype information from TCGA, with the value of "secondary surgery for tumor recurrence: locoregional procedure."

Table 1-1. Data description

Types of data	Platform	Num of Features (*d*)	Num of Records [a] (*N*)	Num of Overlap Records
CNA	Agilent Human Genome CGH Microarray 244A	235,829	278	159
Methylation	Illumina DNA Methylation OMA003 Cancer Panel 1	1,498	235	
Gene Expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043	262	
miRNA	Agilent 8x15K Human miRNA-specific microarray	534	266	

[a]: GBM patients with solid tumor

Methods

I used a graph-based semi-supervised learning as a classification algorithm, which is a halfway learning scheme between supervised and unsupervised learning (41-44). If two patients' samples were more closely related than to others, I assumed that the clinical outcomes of those two patients were more likely to be similar. In other words, clinical outcome prediction can be done by considering relationships between patient samples. A natural method of analyzing relationships between samples is a graph, where nodes depict patient samples and edges represent their possible relations. Figure 1-1 presents a cartoon graph of patient samples. An annotated sample is labeled either by '-1' or '1', indicating the two possible clinical outcomes, either 'normal' or 'cancer.' To predict the label of the unannotated sample '?', the edges connected from/to the sample play an important role in influencing propagation between the sample and its neighbors. This idea can be easily formulated using graph-based SSL (44). Edges represent relations, more specifically similarities between samples that may be extracted from different genomic sources of CNA, methylation, gene expression, miRNA, etc.

Graph-based semi-supervised learning In the graph-based SSL algorithm (44), a sample x_i (i = 1,…,n) is represented as a node i in a graph, and the relationship between samples is represented by an edge. The edge strength from each node j to each other node i is encoded in element w_{ij} of a $n \times n$ symmetric weight matrix W. A Gaussian function of Euclidean distance between samples, with length scale hyperparameter σ, is used to specify

connection strength:

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{(x_i - x_j)^T (x_i - x_j)}{\sigma^2} \right) & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

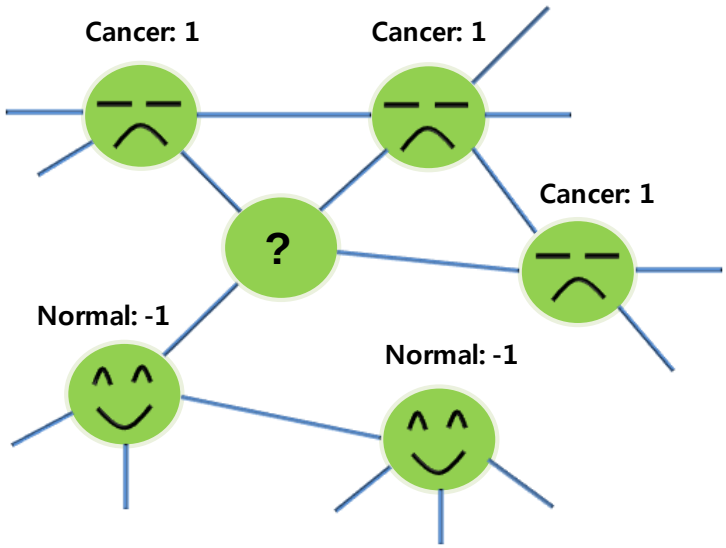


Figure 1-1. A graph model of relationships between patient samples. Nodes represent patient samples and edges depict relations between samples. An annotated sample is labeled either by -1 or +1. In this example, the negative labels indicate samples from 'normal' patients. On the contrary, the positive labels indicate the samples from 'cancer' patients. The clinical outcome of the unannotated sample marked as '?' is predicted by employing graph-based semi-supervised learning.

Nodes *i, j* are connected by an edge if *i* is in *j*'s *k*-nearest-neighborhood or vice versa. Therefore, nearby samples in Euclidean spaces are assigned large edge weights. The labeled nodes have labels $y_l \in \{-1, 1\}$, while the unlabeled nodes have zeros $y_u = 0$. SSL will output an *n*-dimensional real-valued vector $f = [f_l^T f_u^T]^T = (f_1,\ldots,f_l, f_{l+1},\ldots,f_{n=l+u})^T$, which can be thresholded to make label predictions on $f_l=_1,\ldots,f_n$ after learning. It is assumed that f_i should be close to the given label y_i in labeled nodes (loss condition), and overall, f_i should not be too different from the f_i of adjacent nodes (smoothness condition). One can obtain *f* by minimizing the following quadratic functional (41, 43-44):

$$\min_f (f - y)^T (f - y) + \mu f^T L f \qquad (2)$$

where $y=(y_1,\ldots,y_l, 0,\ldots0)^T$, and the matrix *L*, called the graph Laplacian matrix (45), is defined as $L = D - W$ where $D = \mathrm{diag}(d_i)$, $d_i = \sum_j w_{ij}$. The parameter μ trades off loss versus smoothness. The solution of this problem is obtained as

$$f = (I + \mu L)^{-1} y \qquad (3)$$

where *I* is the identity matrix.

Experimental Setting

For each type of genomic data, the five-fold cross-validation (5 CV) was conducted and the performance was measured using the area under the curve (AUC) of receiver operating characteristic (46). The values of model parameters, k and μ, are determined by the results of search over $k \in \{3, 4, 5, 6, 7, 8, 9, 10, 20, 30\}$ and $\mu \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 1.0, 10, 100, 1000\}$. The best combination of model parameters is selected when the best AUC is observed. Figure 1-2 depicts the changes in the AUC over the model parameter variation in the case of CNA (with the 495 features). The best AUC, 0.7102, is achieved when k is 15 and μ is 0.001.

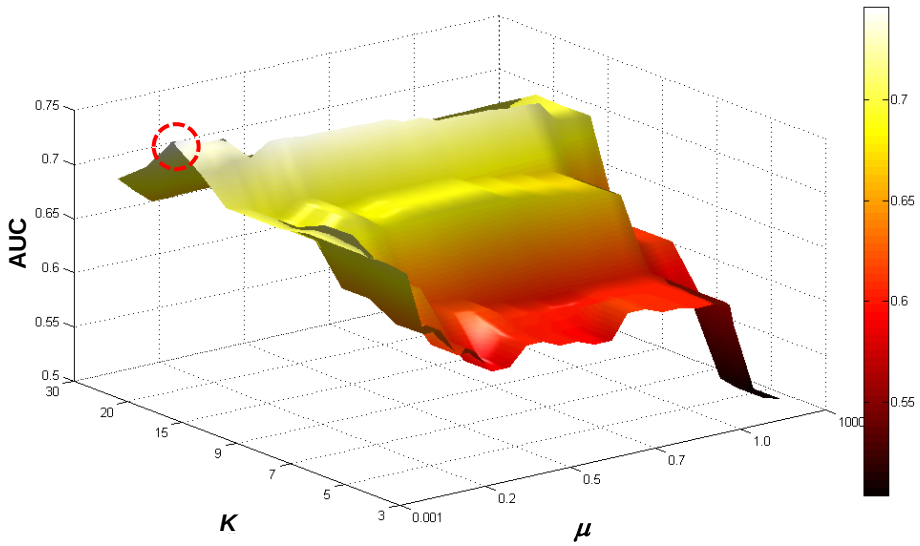


Figure 1−2. **Example of model parameter selection (SSL with CNA data)**

RESULTS

In order to investigate the effect of feature selection, I calculated AUC changes with different number of features as shown in Figure 1-3. In the case of CNA data set, the AUC with 235,829 (p-value < 1) features was 0.4345 while it becomes 0.8131 with the selected 23 features (p-value < 0.001). This validates effect of feature selection. Table 1-2 and Figure 1-4 show comparison results of 5-CV avg. AUC (± standard deviation) for the four data sets. The model parameters are shown on top of the bar in the figure. Among them, the CNA data showed the best performance. The Wilcoxon signed-rank test on CNA versus other data shows the significance of the difference in performance (47).

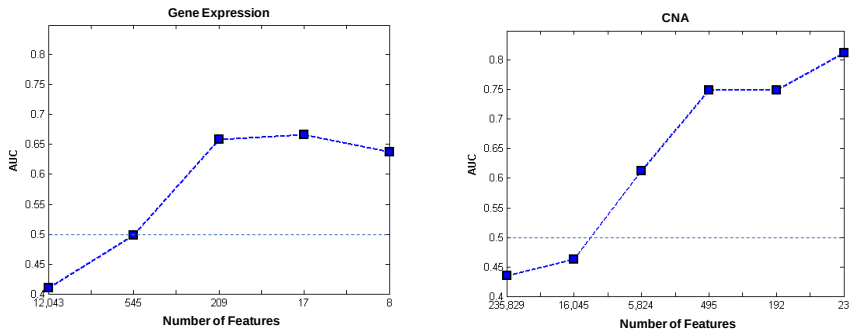


Figure 1-3. **AUC changes with different number of features.** At the *x*-axis the number of features is represented, while *y*-axis shows AUC. Each point of *x*-axis represents the significance level, respectively. (p-value < 1, 0.1, 0.05, 0.01, 0.005, 0.001)

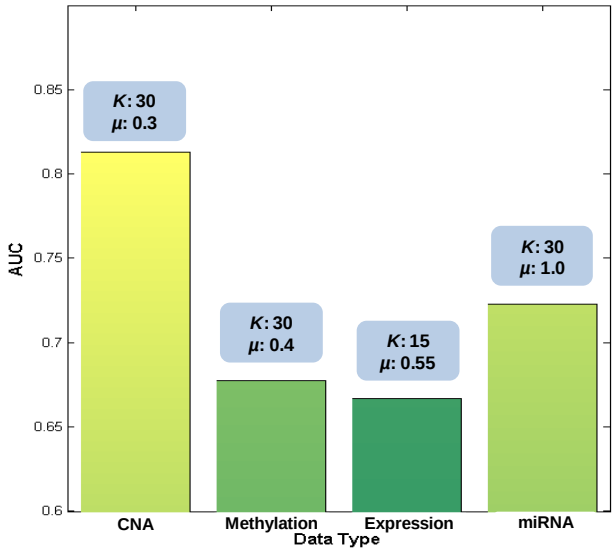


Figure 1−4. **Best AUC comparison between heterogeneous genomic data**

The results showed that the performance of CNA data on the basis of their pattern of chromosomal gains and losses was the best in classification of GBM subtypes. Genomic imbalances including CNAs are hallmarks of various human cancers (8). Human cancer is caused partially by irreversible structural mutations. These can produce alterations in DNA copy number at distinct loci in the genome (48). In contrast to CNA data, gene expression data showed the lowest performance among data set. Several studies have provided insights into the importance of specific CNA in development of solid tumor, showing that these CNA may lead to the altered expression of cancer related genes. However, a small proportion of the genes have been altered simultaneously with gene amplification or deletion in tumors and cancer cell

lines.

Table 1-2. Comparison of significance of the performance between different types of data set

Types of data	Avg AUC	p-value[a]
CNA	**0.7498 (± 0.0241)**	
Methylation	0.5722 (± 0.0437)	0.00
Gene Expression	0.6098 (± 0.0281)	0.00
miRNA	0.5900 (± 0.0427)	0.00

[a]: p-value of Wilcoxon signed-rank test on CNA vs. other dataset

DISCUSSION

In the chapter 1, classification of initial/recurrent tumor in GBM was performed as a base task in order to provide a preliminary insight on the question that is which genomic data is more informative in clinical outcome prediction when multiple genomic dataset are available. Among heterogeneous genomic dataset, CNA data showed the best performance. Thus, for the distinction of these subtypes of GBM, CNA profiling appears to be more advantageous than other different types of genomic dataset. This suggests that CNA data could be solely used in classification of initial vs. recurrent tumor in GBM where other genomic data are unavailable due to the high cost associated with the experimental procedure and sample availability. Although the performance of CNA data was the best, it is still important to identify patterns of other types of genomic data. Because recurrence in GBM is a complex event where many gene products and the other molecules are involved as participants, an empirical comparison on heterogeneous genomic data will be facilitated in revealing molecular mechanisms of recurrence in GBM. Even though this study is limited to the classification problem of initial vs. recurrent tumor in GBM, the comparison framework can be applied to other cancer types or other clinical outcomes such as grade, stage, metastasis, *etc*.

CHAPTER 2

Synergistic effect of different levels of genomic data for cancer clinical outcome prediction

INTRODUCTION

There have been many attempts at cancer classification using a set of miRNA, copy number alterations (CNA), and DNA methylation (7-11). Despite these efforts, however, it still remains difficult to elucidate the development of cancer phenotypes because the cancer genome is neither simple nor independent but rather complicated and dysregulated by multiple molecular mechanisms (12-13). Therefore, no single level of genomic data will be sufficient to comprise all of the information in the mechanism, and hence, a consideration of the layered process of biological systems through incorporation of multiple levels of genomic data will provide a much more reasonable prediction of cancer phenotypes.

In this study, I propose an integrated framework that uses multi-level genomic data sources for the molecular-based classification of clinical outcomes in brain cancer (glioblastoma multiforme; GBM) and ovarian cancer (serous cystadenocarcinoma; OV) (Fig. 2-1). GBM is the most common and aggressive primary brain tumor in adults (38), and notorious for its tendency to recur (39). Despite recent advances in the molecular pathology of GBM, the underling molecular mechanisms associated with clinical outcome are still poorly understood (38, 40). OV is one of the most common gynecological malignancies, and is the 5th leading cause of cancer mortality in women in the United States (49). Understanding the molecular pathogenesis and underlying biology for both types of cancer is expected to provide guidance for improved prognostic indicators and effective therapies.

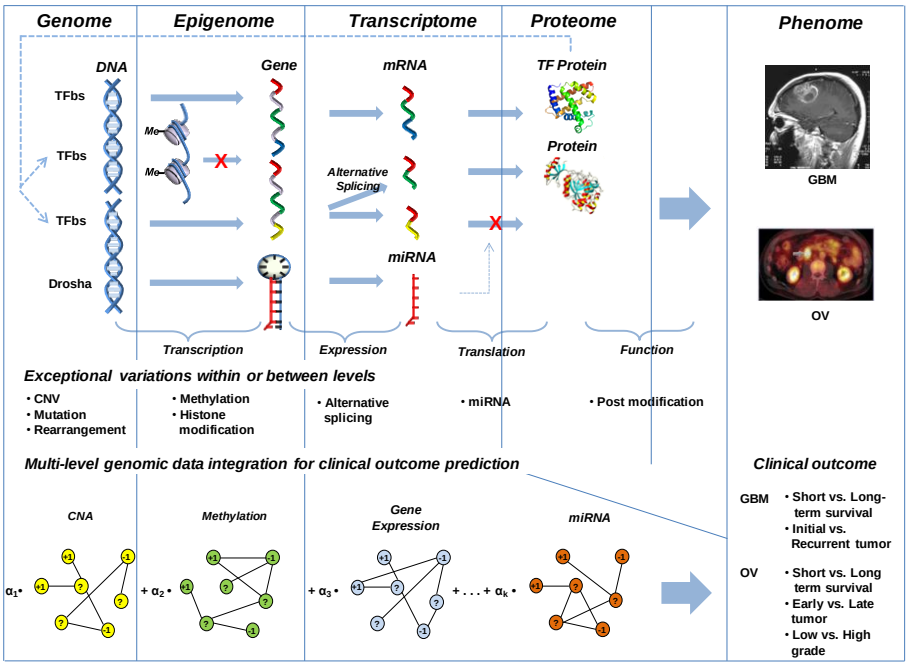


Figure 2−1. **Multi-layers of genomic data in biological system from genome, epigenome, transcriptome and proteome to phenome.** There are many genomic variations within or between levels: Copy number variant (CNV), sequence mutation, and genomic rearrangement in genome level; DNA methylation and histone modification in epigenome level; alternative splicing and miRNA regulation in transcriptome level; post translational modification in proteome level. Multiple graphs given from different genomic levels are integrated into one by finding an optimum value of the linear combination coefficient α_k for the individual graph. TF, transcription factor; TFbs, transcription factor binding site; Me, methylation.

MATERIALS AND METHODS

Data

Datasets were retrieved from the TCGA data portal (http://tcga-data.nci.nih.gov/). Table 2-1 shows the data description of the multi-level genomic datasets in GBM and OV. CNA belongs to the genome level, methylation to the epigenome level, gene expression and miRNA to the level of the transcriptome. The fourth column shows the number of features for each type of genomic data. Five sets of binary classification problems were set using the phenotype information from patients depending on the types of clinical outcomes (Table 2-2). Using the clinical outcome from GBM, the two sets of problems are defined: (1) *short-term or long-term survival* and (2) *initial or recurrent tumor*. In the classification of *short-term or long-term survival*, 'short-term' represents the samples from patients who survived less than nine months, whereas 'long-term' means samples derived from patients who survived longer than 24 months (50). In the classification of *initial or recurrent tumor*, 'initial tumor' indicates samples from surgical resections with no pretreatment history, while samples from secondary surgeries for tumor recurrence are defined as 'recurrent tumor.' Similarly, the remaining three sets of classifications are defined using OV clinical outcomes, which are as follows: (3) *early stage (T1-T2) or late stage (T3-T4)*, (4) *low grade (G1-G2) or high grade (G3-G4)*, and (5) *short-term (< 3 y) or long-term (≥3 y) survival* (51). The last column of Table 2-2 summarizes the number of available (positive/negative) samples for each of these problems.

Table 2-1. Data description

Cancer type	Data type	Platform	# Features (*d*)
GBM	CNA	Agilent Human Genome CGH Microarray 244A	235,829
	Methylation	Illumina DNA Methylation OMA003 Cancer Panel 1	1,498
	Gene expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043
	miRNA	Agilent 8x15K Human miRNA-specific microarray	534
OV	CNA	Agilent SurePrint G3 Human CGH Microarray Kit 1x1M	962,434
	Methylation	Infinium humanmethylation27 BeadChip	27,578
	Gene expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043
	miRNA	Agilent Human miRNA Microarray Rel12.0	799

Table 2-2. Clinical outcomes

Cancer type	Clinical outcome	# Samples (*n*)* (Neg/Pos)
GBM	Short-term survival (survived less than nine months) vs. long-term survival (survived longer than 24 months)	82 (54/28)
	Initial tumor (initial diagnosis) vs. recurrent tumor (tumor recurrence)	159 (39/120)
OV	Short-term survival (survived less than three years) vs. long-term survival (survived longer than three years)	348 (150/198)
	Early stage (T1 or T2) vs. late stage (T3 or T4)	503 (39/464)
	Low grade (G1 or G2) vs. high grade (G3 or G4)	496 (65/431)

*Solid tumor samples from each type of cancer were only considered.

Clinical Outcome Classification

Graph-based semi-supervised learning I used a graph-based semi-supervised learning as a classification algorithm, which is a halfway learning scheme between supervised and unsupervised learning (41-44). Edges represent relations, more specifically similarities between samples that may be extracted from different genomic sources of CNA, methylation, gene expression, miRNA, etc. Different data sources produce different graphs. However, clinical outcome prediction can benefit by integrating diverse graphs from diverse genomic data sources, rather than relying only on single sources that may have possible limitations, (i.e. incomplete information and noise). When data sources are presented as a graph form, combining multiple data sources can be done by employing a graph integration method (32-34).

Integration of Multi-level Genomic Data Sources From multi-level genomic data sources, multiple graphs are generated. Information from each graph is regarded as partially independent from and partly complementary to others. Therefore, it is not accurate enough to elucidate phenotype using only a single genomic data source belonging to a specific single layer. Reliability may be enhanced by integrating all available information sources using the method proposed by Tsuda *et al*. (2005), which has been re-validated on the extended problem of protein function prediction (32). According to the method, the integration of multiple graphs is used to find an optimum value of the linear combination coefficient for the individual graphs. This corresponds to finding

the combination coefficients α for the individual Laplacians of the following mathematical formulation:

$$\min_{\alpha} y^{\mathrm{T}}(I+\sum_{k=1}^{K}\alpha_k L_k)^{-1}y, \qquad \sum_k \alpha_k \le \mu \tag{1}$$

where K is the number of graphs (data sources) and L_k is the corresponding graph-Laplacian of graph G_k. One can perceive the formulation by synchronizing the schematic descriptions shown in Figure 2-1. Similar to the output prediction for single graphs, the solution is obtained by

$$f = (I+\sum_{k=1}^{K}\alpha_k L_k)^{-1}y. \tag{2}$$

RESULTS

From the multi-levels of genomic data available from TCGA, the following five sets of clinical outcome classifications were defined: for GBM, (1) short-term or long-term survival and (2) initial or recurrent tumor and for OV (3) early or late stage; (4) low or high grade; (5) short-term or long-term survival. For the five sets of binary classification problems, the proposed approach, *prediction from integration of multi-level genomic data sources*, was compared with the four *individual predictions* obtained from CNA, methylation, gene expression, and miRNA, respectively. For each problem, I calculated the five-fold cross-validation (5 CV) area under the curve (AUC) of receiver operating characteristic (ROC) (46) and the true positive rate yielding an 1% false positive rate (TP1FP) as performance measurements (28). The Wilcoxon signed-rank test was used to validate the significance of the difference in performance for all the combinations of the comparisons (47).

Glioblastoma Multiforme

Table 2-3 shows the AUC performance on the two sets of classifications of GBM clinical outcomes. The AUCs of the four *individual sources* (CNA, methylation, gene expression, and miRNA) are shown in the first four rows and the AUC of the proposed approach—*integration with multi-level genomic data sources* is shown in the fifth row. For the short-term survival vs. long-term survival classification, the SSL with gene expression data performed best with 0.8560 AUC (underlined) among the four genomic data sources, and

CNA data showed comparable performance with an AUC of 0.8160. However, none of the AUC values from single data sources could outperform the AUC of 0.8760 (boldface) generated by the multi-level genomic data. The p-values of the pairwise comparisons in AUC between the proposed approach and the single data approaches demonstrate that a statistically significant difference exists in performance. The superior performance of the integration approach is also found in terms of the value of the TP1FP— of 0.8, the highest among the five. Furthermore, for the initial tumor vs. recurrent tumor classification, the SSL with CNA data showed the best performance with an AUC value of 0.8131 (underlined) among the four individual data sources, but again the performance of the integration approach was superior to that of the best individual approach with an AUC of 0.8369 and TP1FP of 0.75 (boldface) was superior to all of the individual approaches.

Table 2-3. AUC results on GBM clinical outcomes

Clinical outcome	Data type	AUC (*P-value**)	TP1FP
Short-term survival vs. long-term survival	CNA	0.8160 (2.19e-26)	0.30
	Methylation	0.7408 (1.19e-28)	0.60
	Gene expression	<u>0.8560 (1.22e-11)</u>	0.72
	miRNA	0.7480 (1.07e-28)	0.40
	Multi-level data	**0.8760**	**0.80**
Initial tumor vs. recurrent tumor	CNA	<u>0.8131 (3.04e-04)</u>	0.65
	Methylation	0.6774 (3.30e-33)	0.20
	Gene expression	0.6667 (2.09e-34)	0.15
	miRNA	0.7226 (1.15e-33)	0.43
	Multi-level data	**0.8369**	**0.75**

*The p-value of the pairwise Wilcoxon signed-rank test in AUCs between the multi-level integration approach and the single data approaches

Serous Cystadenocarcinoma

Table 2-4 shows the AUC performance of the three sets of classification problems of OV clinical outcomes. For the short-term survival vs. long-term survival classification, the SSL with gene expression data performed best with an AUC of 0.7651 when compared with the other single genomic data sources. Note that a similar result was obtained for GBM, which might imply that the information from gene expression data plays a critical role in the classification of short-term vs. long-term survival. Among the four individual sources, the best performing data differed for each classification; for the early stage vs. late stage classification, the prediction from CNA data showed the best performance (0.8767 AUC) while for the low grade vs. high grade classification, methylation data performed best (0.8161 AUC). However, the AUC of the proposed approach consistently outperforms the individual best for all of the three sets of problems, attaining values of 0.7867, 0.8932, and 0.8678. There is a slight degradation of the proposed approach in TP1FP for the low grade vs. high grade classification, but the magnitude in difference is negligible (0.57 in methylation data vs. 0.54 in multi-level integrated data).

Table 2-4. AUC results on OV clinical outcomes

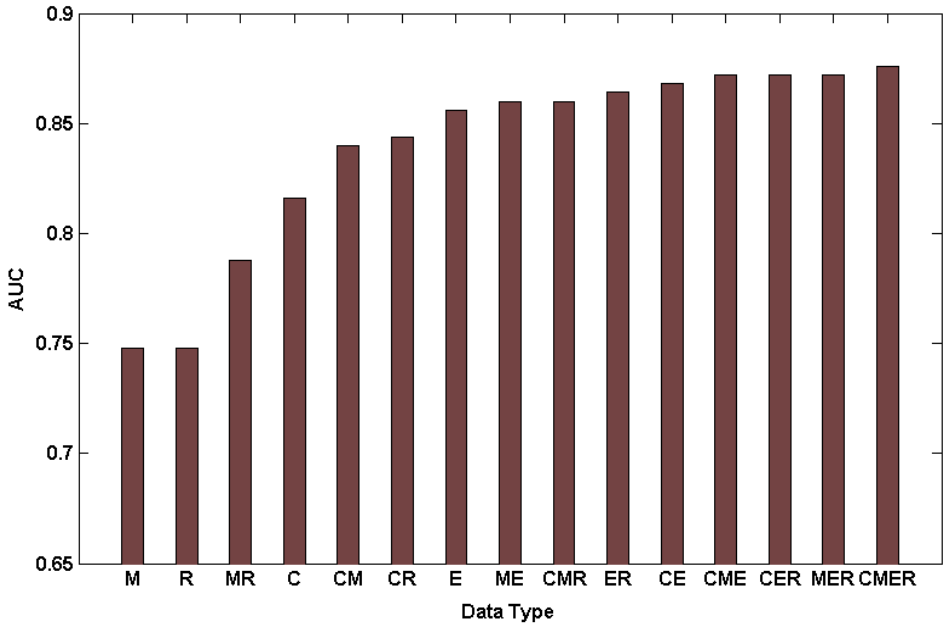
Clinical outcome	Data type	AUC (*P-value**)	TP1FP
Short-term survival vs. long-term survival	CNA	0.6547 (1.24e-28)	0.17
	Methylation	0.7251 (1.34e-27)	0.14
	Gene expression	0.7651 (8.96e-10)	0.26
	miRNA	0.6403 (1.24e-28)	0.17
	Multi-level data	**0.7867**	**0.40**
Early stage vs. late stage	CNA	0.8767 (1.87e-05)	0.74
	Methylation	0.7149 (1.51e-28)	0.61
	Gene expression	0.8332 (2.31e-05)	0.53
	miRNA	0.7661 (1.39e-21)	0.78
	Multi-level data	**0.8932**	**0.80**
Low grade vs. high grade	CNA	0.8014 (3.43e-05)	0.37
	Methylation	0.8161 (4.63e-09)	**0.57**
	Gene expression	0.7676 (2.59e-06)	0.39
	miRNA	0.6887 (9.61e-15)	0.16
	Multi-level data	**0.8678**	0.54

*The p-value of the pairwise Wilcoxon signed-rank test in AUCs between the multi-level integration approach and the single data approaches

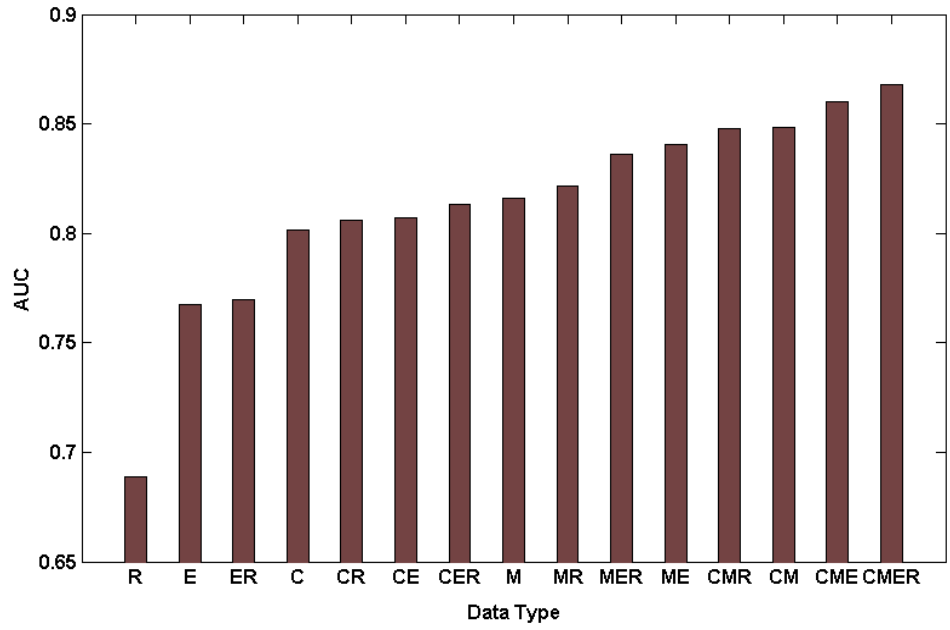
Integration Effect

I found that the integration of various genomic data sources increases the performance of clinical outcome prediction. As a next step, I investigated the effect of integration by considering all possible combinations of the four multi-level genomic data sources. Figure 2-2 shows a gradual increase in AUC by integration: for the short-term survival vs. long-term survival classification of GBM (Fig. 2-2A) and for the low grade vs. high grade classification of OV (Fig. 2-2B), where C stands for CNA, M for methylation, E for gene expression, and R for miRNA, and the combinations are represented as MR, CMR, and so on. AUC consistently increases as more data sources are added to the combination in both GBM and OV. In Fig. 2-2A, for instance, AUC increases in the order of mass of combinations, C < CR <

CMR < CMER. These findings suggest that biological information may be fused to various data sources from different genomic levels; therefore, integration of those independent or complementary pieces of information may elevate the opportunity of success in prediction of clinical outcomes in cancer.



(A) GBM: short-term survival vs. long-term survival



(B) OV: low grade vs. high grade

Figure 2−2. **Gradual increase in AUC by integration: C stands for CNA, M for methylation, E for gene expression, and R for miRNA, and the combinations are represented as MR, CMR, and so on.** (A) The short-

term vs. long-term survival classification of GBM. (B) The low vs. high grade classification of OV.

DISCUSSION

In the chapter 2, a pilot framework of integration of multiple levels of genomic data sources, CNA, DNA methylation, gene expression, and miRNA expression, has been applied to the problem of prediction of clinical outcomes in GBM and OV. On the basis of the results of our computational experiments, some biological and clinical implications may be cautiously drawn. Figure 8 illustrates the following observations that show the level of contributions of multi-level genomic data sources for the five classification problems. As of yet, there has been no clear-cut definition on boundaries between the different genomic levels; but, it is naturally conceived that the *structural changes* in the chromosome or chromatin will lead to the changes on data sources obtained from the genome or the epigenome level (CNA and methylation in the experiment) before the influence reaches to the transcriptome level. On the other hand, the *functional changes* caused by the by-products of DNA will be more directly related to the changes on data sources generated from the transcriptome level (mRNA and miRNA in our experiment).

First, the CNA data performed best in the initial vs. recurrent tumor classification in GBM and the early vs. late stage classification in OV. Both problems concern the structural changes in chromosome by the elapsed amount of time since tumor initiation (52-53). Therefore CNA data might have provided appropriate information for classifying the alternative clinical outcomes.

Second, the performance of the gene expression data was superior to those of others in the short-term vs. long-term survival classification in both GBM and OV. The strength of the current malignant behavior of the tumor is related to the functional changes of genes or proteins (6) which can be detected by gene expression data in our experimental setting. Interestingly, the gene expression data performed as good as the other three dataset (CMR), and almost as good as the full dataset (CMER) (Fig. 2-2A). This calls for additional bioinformatical analyses. One intriguing possibility suggests that the same genomic loci contribute clinical information in more than one domain – the same genes that change in their copy number and methylation patterns also present predictive powers based on mRNA expression levels.

Third, the methylation data performed best for the low vs. high grade classification of OV. Despite the lack of understanding of epigenomic characteristics in cancer, I suggest that structural changes may be worthy of further study.

Fourth, even though I made an *ad-hoc* separation on genomic data to structural changes or functional changes, the phenotype of clinical outcome is not influenced by only one of them. As shown in Figure 2-3, the integration of all genomic data sources can be helpful to unveil the relationship from genome to phenome.

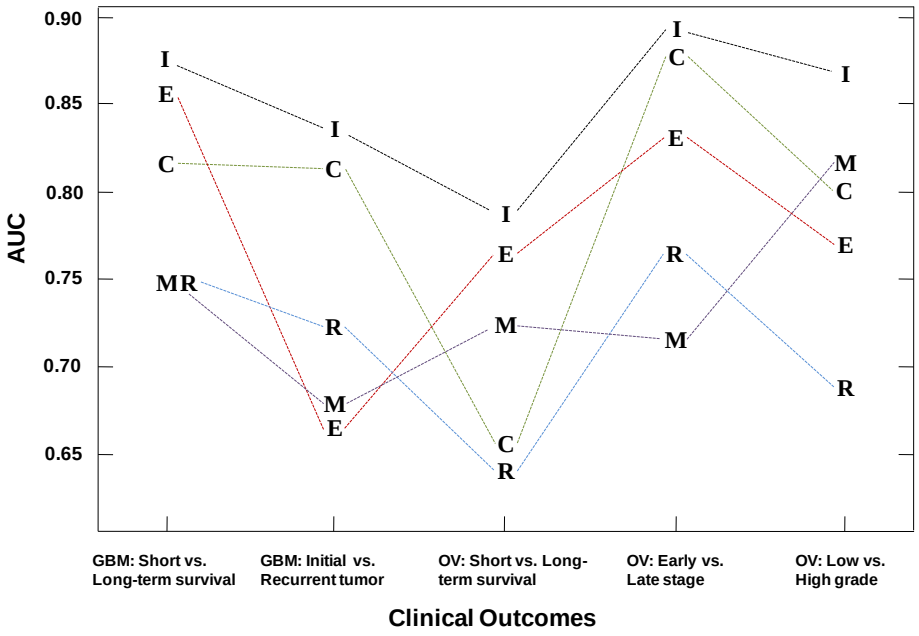


Figure 2−3. **Performance comparison of genomic data over the five sets of clinical outcome classification problem: C stands for CNA, M for methylation, E for gene expression, R for miRNA, and I for the integration of the four data sources.**

CHAPTER 3

Combining multi-layers of genomic data and inter-relationship between different layers of genomic features

INTRODUCTION

The limitation of previous chapter is integration with multi-layers of genomic data for cancer clinical outcome prediction without considering of inter-relationship information between them. There are possible relationships between the sample features (attributes) belonging to different layers of genomic data such as ‘miRNA-target genes,’ ‘copy number alteration region-genes located in the alteration region,’ ‘DNA methylation site-specific genes regulated by promoter regions,’ etc. Therefore, when integrating multiple genomic data, it will be desirable that a framework will be capable of containing the inter-relationships between sample features belonging to different layers of the biological system. From multi-level genomic data sources and inter-relationship information, multiple graphs are generated. Information from each graph is regarded as partially independent from and partly complementary to others. Thus, it is not accurate enough to elucidate phenotype using only a single genomic data source belonging to a specific single layer. This study can be applied to three types of sub-studies.

A. miRNA – mRNA dataset

Expression profiling of mRNA has been used to molecularly characterize various tissues and tumor. However, gene expression regulation through miRNA as one of the major regulators has attracted much attention during recent years. miRNAs are short ribonucleic acid (RNA) molecules, on average only 22 nucleotides long and are found in all eukaryotic cells. miRNAs are

involved in the post-transcriptional regulation of genes either by mRNA cleavage and degradation (Fig. 3-1) or by repressing the translation of mRNA into protein (54-55). Many miRNAs regulate genes associated with different biological processes such as development, proliferation, apoptosis, stress response, and tumorigenesis (56-60). Therefore, I assume that accuracy of prediction model increases because of incorporation of information fused over genomic datasets (mRNA and miRNA) and inter-relationship between them, providing an enhanced global view on miRNA regulation mechanism in cancer. There is a many-to-many relationship between miRNAs and mRNAs since a single miRNA targets multiple mRNAs and a single mRNA is targeted by multiple miRNAs. The miRNA-target gene information will be used from miRecords which is integrated resource for miRNA-target interaction (61).

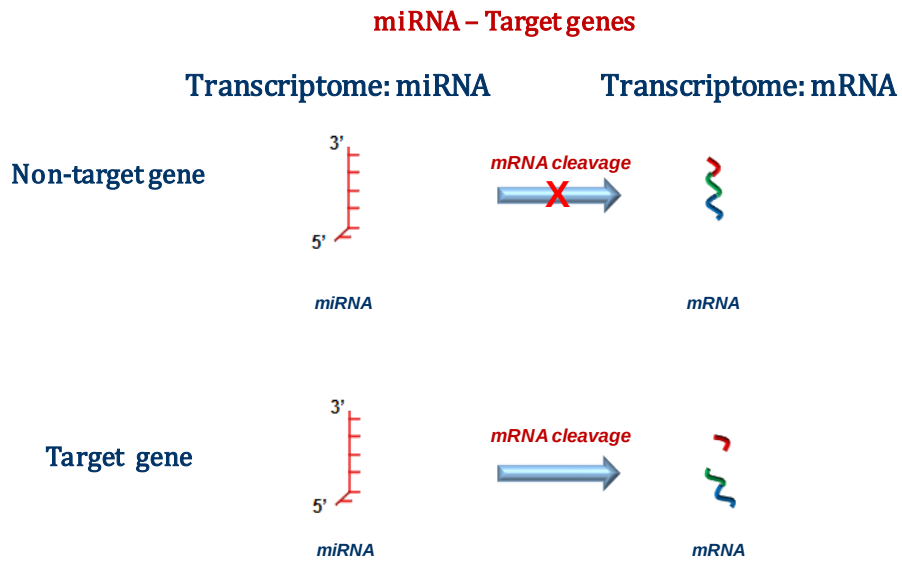


Figure 3-1. **Mechanism of messenger RNA cleavage specified by a miRNA**

B. Copy number alteration – mRNA dataset

Irreversible structural mutations in human cancer can produce changes in DNA copy number at distinct locations in the human genome (48). Copy number alterations affect the function of genes and ultimately cause a different phenotype in cancer. Most tumors show numerous genomic alterations, but it has been a challenge to identify those that are required for different stages of tumor development. As most genome alterations affect the transcriptome, it would be useful to integrate genome profiling with transcriptome profiling. Pollack et al. profiles DNA copy number alterations across 6,691 mapped human genes in 44 samples of predominantly advanced, primary breast tumors and 10 breast cancer cell lines (62). Paired gene expression allowed assessment of the extent to which variation in gene copy number contributes to variation in functional changes in tumor cells (Fig. 3-2). About 62% of highly amplified genes showed increased expression levels. Additionally, DNA copy number correlated with gene expression across a range of DNA copy number alterations, including deletions. On average, a twofold change in DNA copy number was associated with a corresponding 1.5 fold change in mRNA levels.

Thus, when integrating copy number alteration and gene expression data, it will be desirable that a framework will be capable of containing the inter-relationships between copy number alteration regions and genes in the altered region, providing an enhanced global view on functional effect of copy number alterations in cancer.

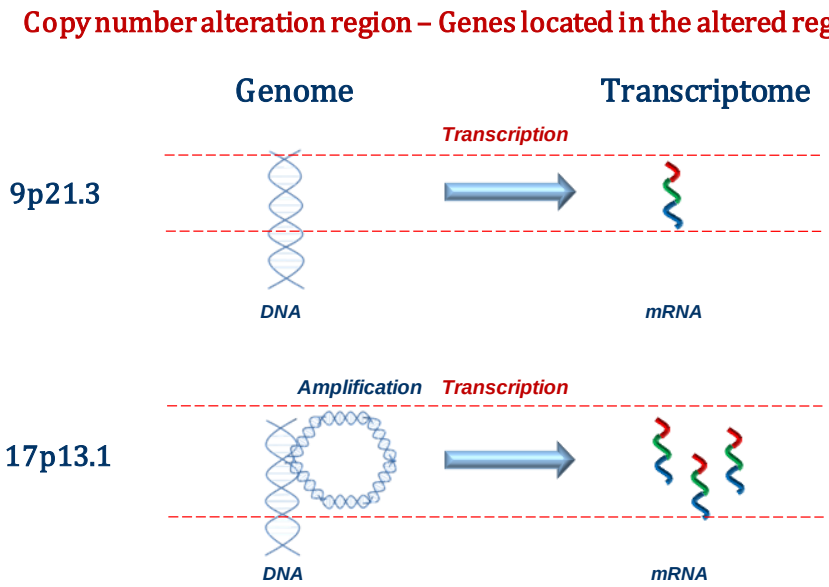


Figure 3−2. **Example of up-regulated genes affected by copy number amplification**

C. DNA methylation – mRNA dataset

DNA methylation plays an essential role in normal development through its effects on gene imprinting, X-chromosome inactivation, and transcriptional silencing of repetitive elements (63). However, genome-wide patterns of DNA methylation are altered with hypermethylation of a subset of CpG islands and hypomethylation of repeat intergenic regions in cancer. In general, cancer cells show hypomethylation patterns of intergenic regions that normally consist of the majority of methyl-cytosine content of a cell. Simultaneously, cancer cells exhibit hypermethylation within the promoter regions of many CpG island-associated tumor suppressor genes. As a result, these regulatory genes are transcriptionally silenced resulting in a loss-of-function (Fig. 3-3).

Therefore, DNA methylation significantly affects the global genomic regions of cancer cells, potentially to an even greater extent than coding region mutations through the effects of both hypomethylation and hypermethylation (64). When integrating DNA methylation and gene expression data, I could get the inter-relationship information that gene up-regulation can be due to hypomethylation (decrease in methylation of cytosine and adenosine residues in DNA) and down-regulation due to hypomethylation. Thus, I assume that combining DNA methylation and gene expression with inter-relationship information between them is better to elucidate the cancer phenotypes.

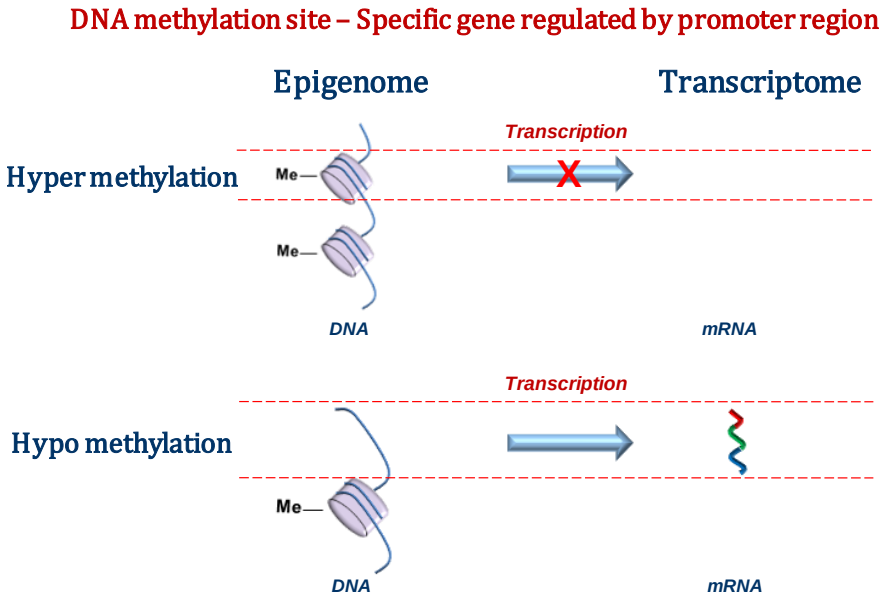


Figure 3-3. **DNA methylation patterns are altered in cancer**

In addition, despite the lack of understanding about other pairs of genomic data, it can be extended to investigate the effect of inter-relationship between following data set (Fig. 3-4).

D. Copy number alteration – miRNA data set

E. Copy number alteration – DNA methylation data set

F. DNA methylation – miRNA data set

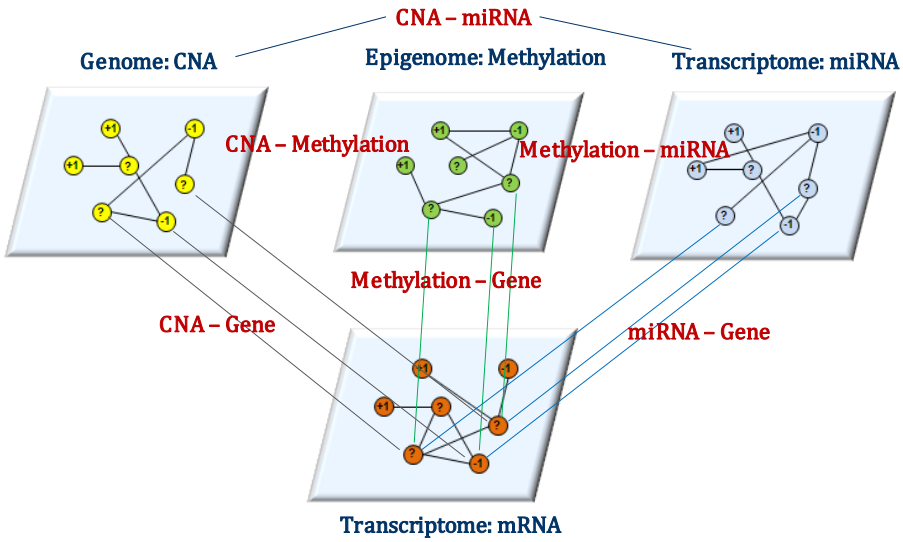


Figure 3-4. **All pairwise inter-relationship between different levels of genomic data**

Improving our understanding of complex mechanisms in cancer and developing analytic models will require an increased understanding of the contributions of and interactions between multi-layers of genomic data that

contribute to tumor formation and progression. Therefore, when integrating multiple genomic data, it will be desirable that a framework will be capable of containing the inter-relationships between sample features belonging to different layers of the biological system (Fig. 3-5).

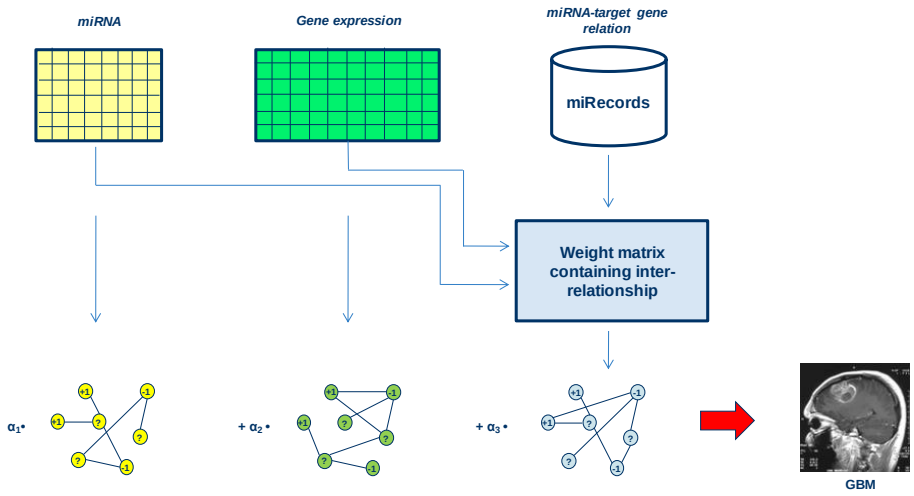


Figure 3-5. **Schematic overview of combining different levels of genomic data and inter-relationship (miRNA – gene expression)**

In order to construct a similarity matrix as an input to graph-based semi-supervised learning, similarity matrices from each genomic dataset and inter-relationship between different layers of genomic dataset can be filled in a 4 x 4 matrix (Fig. 3-6). A similarity matrix from inter-relationship between different layers of genomic dataset could be generated through the previous study, for example, miRNA and gene expression data. A comprehensive framework will be valuable for elucidating cancer phenotype by incorporation

of multiple genomic data sources and inter-relationship information between them.

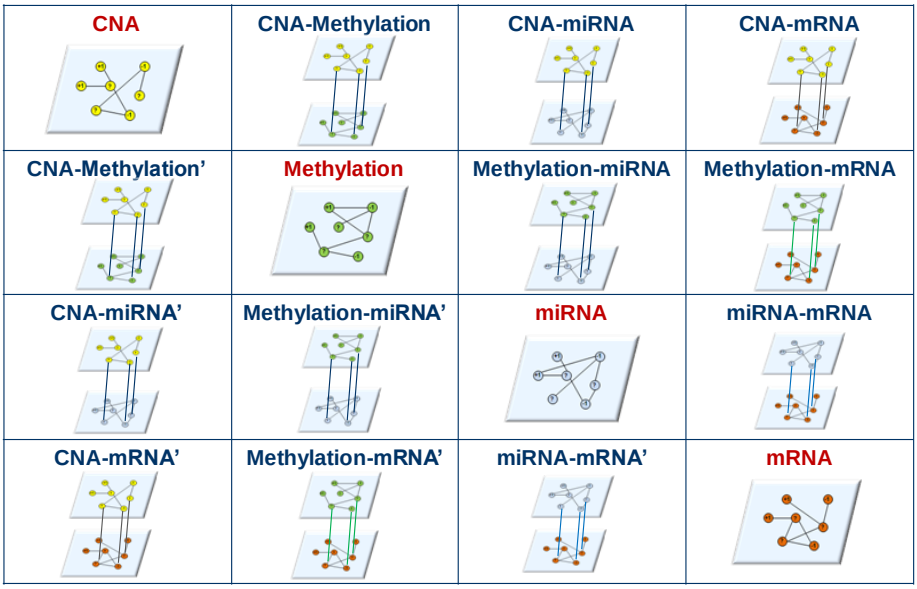


Figure 3-6. **Construction of similarity matrix for inter-relationship between different levels of genomic data**

In order to demonstrate the validity of the effect of integration with inter-relationship, gene expression and miRNA data were used and analyzed as a pilot task. Gene expression profiling has been used to molecularly characterize various tumors and tissues. However, regulation of gene expression by microRNAs (miRNAs) has attracted much attention recently. MicroRNAs are short ribonucleic acid (RNA) molecules, on average only 22 nucleotides long and are found in all eukaryotic cells. MicroRNAs are

involved in the post-transcriptional regulation of genes either by inducing degradation of the transcript of their multiple targets or by repressing the translation of mRNA into protein (54-55). MicroRNAs regulate many genes associated with different biological processes such as development, stress response, apoptosis, proliferation, and tumorigenesis (56-60).

DNA microarrays have already been widely used for the classification of tumor subtypes or clinical outcomes for the diagnosis, treatment, or prognosis of cancer for many years (1-6). Despite these efforts, however, the elucidation of cancer phenotypes remains problematic since the cancer genome is neither simple nor independent but is complicated and dysregulated by multiple mechanisms in the biological system (12-13). Therefore, when elucidating of cancer phenotype using gene expression data, it will be desirable that a framework will be capable of containing the external knowledge such as inter-relation between genomic features belonging to different levels of genomic data.

In computational biology, a novel knowledge has been obtained mostly by identifying ‘intra-relation,’ the relation between entities on a specific biological level such as gene expression or miRNA, and many such researches have been successful. However, intra-relations are not fully explaining complex cancer mechanisms because the information that relations between miRNAs and target genes are strongly associated with different biological processes is missing (65-67). As many target prediction algorithms of miRNAs have been studied recently (61, 68-69), the ‘inter-relation’ between miRNA and gene expression can be constructed from biological experimental

data as well as genomic knowledge. Thus, I assume that accuracy of prediction model increases because of incorporation of information fused over genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression, providing an enhanced global view on miRNA regulation mechanism in cancer. Through this approach, the problem addressed here is how informative inter-relationship between miRNA and gene expression for cancer clinical outcome prediction is.

In this study, I propose an integrated framework that combines genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression for the molecular-based classification of clinical outcomes. In order to demonstrate the validity of the proposed method, the prediction of short-term/long-term survival for 82 patients in glioblastoma multiforme (GBM) is adopted as a base task. GBM is the most common and aggressive primary brain tumor in adults (38), and notorious for its tendency to recur (39). Despite recent advances in the molecular pathology of GBM, the underling molecular mechanisms associated with clinical outcome are still poorly understood (40).

MATERIALS AND METHODS

Data

Datasets were retrieved from the Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/). A binary classification problem was set using the phenotype information from patients depending on the survival information. In the classification of *short-term or long-term survival*, 'long-term' means samples derived from patients who survived longer than 24 months (50). The total 82 patients' records were available across the miRNA and gene expression data sets (N=82), in which 54 were short-term survival while the remaining were long-term survival.

Retrieving mRNA targets of miRNA

There is a many-to-many relationship between miRNAs and mRNAs since a single miRNA targets multiple mRNAs or a single mRNA is targeted by multiple miRNAs. In order to get target relations between miRNA and mRNA, I used miRecords which is integrated resources of miRNA that store target interactions produced by 11 established miRNA target prediction programs (61). Among 11 algorithms, a binary relation between miRNA and mRNA was set when more than 3 algorithms provide the target relation.

Prediction based on intra-relation among mRNAs

I used a graph-based semi-supervised learning (SSL) as a classification algorithm, which is a halfway learning scheme between supervised and

unsupervised learning (41-44). If two patients' samples were more closely related than to others, I assumed that the clinical outcomes of those two patients were more likely to be similar. Thus, clinical outcome prediction can be done by considering similarities between patient samples.

Prediction based on inter-relationship from miRNA to mRNA

The main problem of this study is to develop an adequate measure to calculate the similarity matrix containing inter-relationship information between miRNA and gene expression. There are many measures to construct the similarity matrix for graph-based semi-supervised learning such as *k*-NN graphs, ε-NN graphs, tanh-weighted graphs, exp-weighted graphs, etc (42). For these methods, there is an assumption that the length of vector from two matrices or matrix itself should be same in order to calculate the similarity. However, it is difficult to calculate the similarity matrix containing inter-relationship information between miRNA and target genes because the length of vector from two matrices is different, for example 534 miRNAs and 12,043 genes in miRNA and gene expression, respectively (Fig. 3-7 (A)). Thus, a new measure for calculating the similarity matrix containing inter-relationship information from different levels of genomic data has been developed in this study (Fig. 3-7 (B)).

MicroRNA dataset is represented by i patients ($i = 1,\ldots,N$) and l miRNAs ($l = 1,\ldots,N_{mi}$) and gene expression dataset is represented by j patients ($j = 1,\ldots,N$) and m genes ($m = 1,\ldots,N_G$) (Fig. 3-7 (A)). The edge strength from each miRNA patient to each gene expression patient is encoded in element w_{ij}

of an *N*x*N* weight matrix. A weight matrix containing inter-relationship information between miRNA and target genes is obtained by

$$f_{ij} = \sum_{l=1}^{N_{mi}} \sum_{m=1}^{N_G} miRNA(i,l) \bullet gene(j,m) \quad (1)$$

where *m*-th gene is targeted by *l*-th miRNA. After calculating f_{ij}, each element is normalized and transformed by

$$Z_{ij} = \frac{f_{ij} - \overline{f}}{std(f)} \quad (2)$$

$$w_{ij} = \frac{1}{1 + e^{-Z_{ij}}} \quad (3)$$

Integration of multiple graphs In order to combine the graph from gene expression and the reconstructed graph via inter-relationship, two graphs can be integrated from finding optimum combination coefficients. Information from each graph is regarded as partially independent from and partly complementary to others. Reliability may be enhanced by integrating all available data sources using the method proposed by Tsuda *et al*. (2005), which has been re-validated on the extended problem of protein function prediction (32).

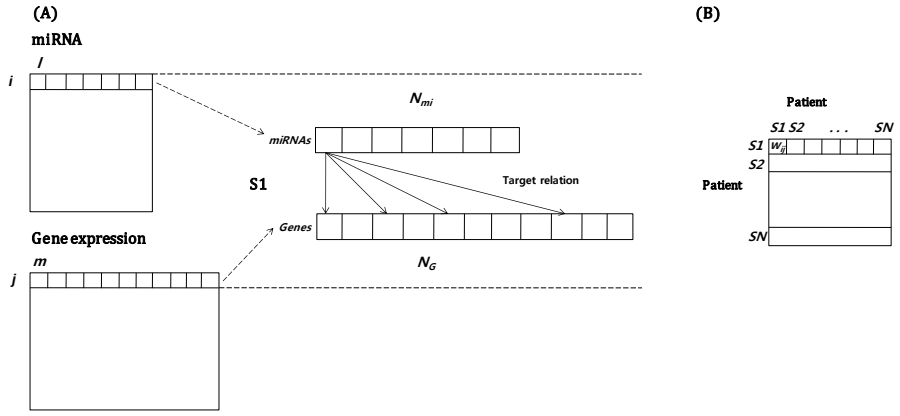


Figure 3-7. **Graphical data description** (A) Data structure of miRNA, gene expression and their target relation (B) Similarity matrix containing inter-relation between miRNA and gene expression

Experimental Setting

In order to evaluate the effect of inter-relation between miRNA and target genes, the intra-relation of gene expression was reconstructed from inter-relation between miRNA and gene expression. I defined the 4 cases of graph for demonstrating the validity of the proposed method (Fig 3-8).

(A) Original graph from gene expression (G_O): I made an original graph from gene expression data where nodes depict patients and edges represent their possible relations.

(B) Damaged graph from the original graph (G_D): I randomly reduced the edges from the original graph, G_O, in order to make the incomplete graph. G_{D50} means the gene expression graph with 50 percent of damaged edges.

(C) Reconstructed graph via inter-relationship (G_R): Reconstructed graph of gene expression was generated via inter-relationship between miRNA and gene expression.

(D) Augmented graph (G_A): An augmented graph was generated by combining damaged graph (G_D) from the original graph and reconstructed graph (G_R) from inter-relation.

Since genomic data sources are generally high dimensional and noisy, and contain many redundant features, which may incur computational difficulty and low accuracy, a Student *t*-test based feature selection method was used (70). Even though there are many feature selection techniques such as filter, wrapper, and embedded method (71), a simple univariate feature selection method was used in order to emphasize not the effect of feature selection but the effect of integration with inter-relationship between miRNAs and target mRNAs.

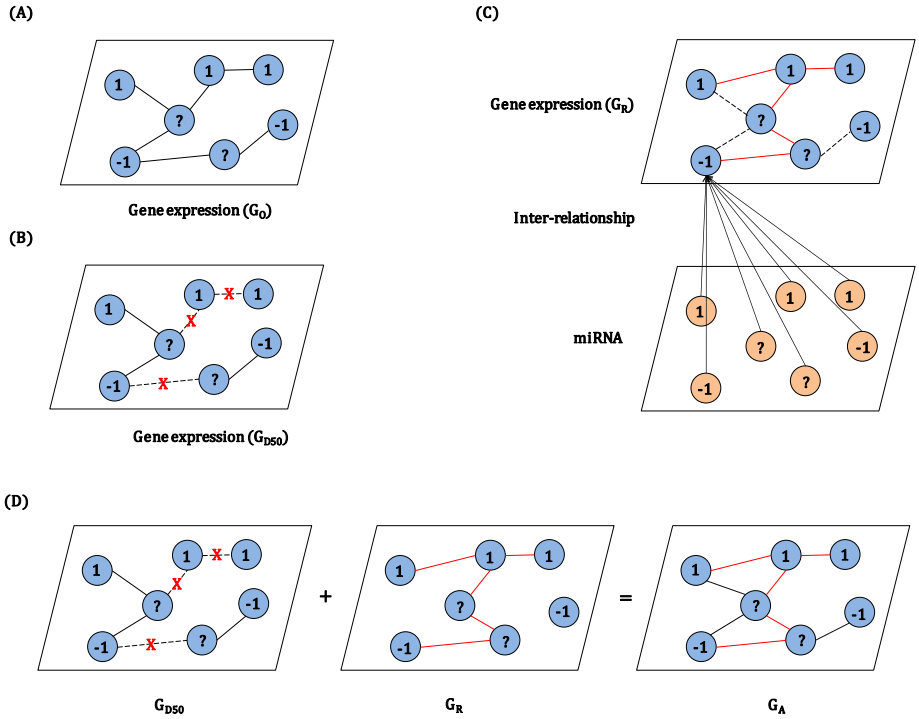


Figure 3-8. **Example model of the original, damaged, reconstructed, and augmented graphs** (A) G_O: Original graph from gene expression (B) G_{D50}: Gene expression graph with 50 percent of damaged edges (C) G_R: Reconstructed graph via inter-relationship between miRNA and gene expression. Red lines represent edges from inter-relation and dashed lines shows the edges from the original graph. (D) G_A: Augmented graph by combining 50 percent of damaged graph and reconstructed graph

RESULTS

The receiver operating characteristic (ROC) curve plots sensitivity (true positive rate) as a function of 1-specificity (false positive rate) for a binary classifier system as its discrimination threshold is varied (46). An ROC score of 0.5 corresponds to random prediction, and an ROC score of 1.0 implies that the model succeeded in putting all of the positive examples before all of the negatives. For each problem, I calculated area under the curve (AUC) of ROC as a performance measure. Each experiment is repeated three times in order to estimate the variance of the measurement values and five-fold cross-validation was conducted. The Wilcoxon signed-rank test was used to assess the significance level of difference in performance between the results of damaged graphs and augmented graphs (47).

Experimental results

Figure 3-9 shows the prediction performance on the classification of short-term and long-term survival for 4 cases of proposed graphs. The AUCs of the 4 graphs (original graph from gene expression data (G_O), damaged graph from the original one (G_D), reconstructed graph via inter-relation between miRNA and mRNA (G_R), and augmented graph by damaged graph and reconstructed graph (G_A)) are shown in the *y* axis and the percent of damaged edges are represented in the *x* axis. The main result of our study is that the prediction performance was improved by integrating the original gene expression (G_O) and the reconstructed graph via inter-relation between miRNA and mRNA

(G_R) (Fig 3-9). I found that the opportunity for success in prediction of clinical outcomes in GBM was increased when the prediction was based on the integration of genomic data and genomic knowledge based on inter-relationship.

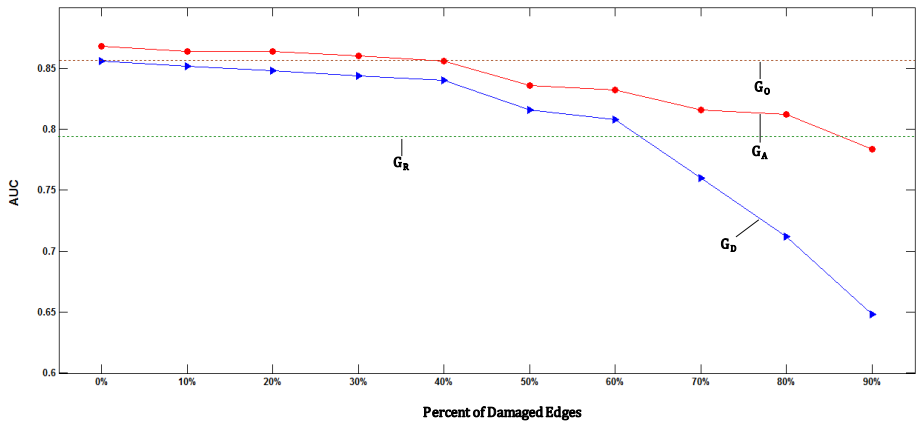


Figure 3-9. **Performance comparison of 4 cases of graphs** G_O: Original graph from gene expression (brown-dashed line), G_D: Gene expression graph with damages (blue line), G_R: Reconstructed graph via inter-relationship between miRNA and mRNA (dark green-dashed line), G_A: Augmented graph by damaged graph and reconstructed graph (red line)

As the percent of damaged edges in gene expression graph increased, the AUCs of damaged graph (G_D) are getting decreased sharply compared to the original graph from gene expression data (G_O) (Fig 3-9). However, the performances of the augmented graph (G_A) showed robust results even though 90 percent of edges were reduced from the original graph. The performance of G_A, a graph combining biological experimental data and genomic knowledge,

is higher than the one of G_O, an original graph from gene expression only, from 0 to 30 percent of damaged edges (Fig 3-9). This suggests that genomic knowledge is complementary to the prediction power of explaining cancer phenotype even though biological experimental data such as gene expression has incomplete information.

The significance level of difference in performance between the results of damaged graph and augmented graph was conducted using Wilcoxon signed-rank test (Table 3-1). The level of significance increased as long as the percentage of damaged edges increased. Figure 3-10 shows a gradual increase in AUC by augmented graph. Dark blue bar represents the results from damaged graph and brown bar depicts the one from augmented graph. Light blue bars indicate the AUC of the original graph and reconstructed graph, respectively. This provides improving performance from the augmented knowledge based on inter-relation between mRNA and miRNA.

Table 3-1. Significance test of the performances between G_D and G_A

Percent of damaged edges	AUC of G_D	AUC of G_A	P-value
10%	0.852	0.864	1.80e-03
30%	0.844	0.860	2.10e-03
50%	0.816	0.836	1.91e-04
70%	0.760	0.816	2.38e-04
90%	0.648	0.784	2.36e-05

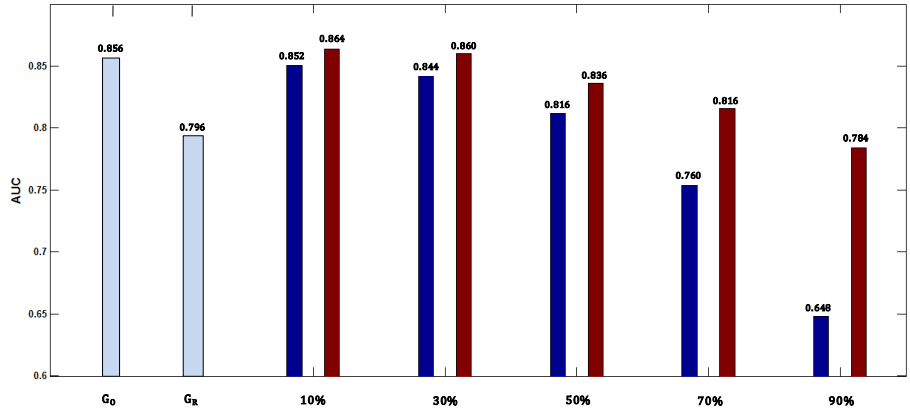


Figure 3-10. **Improving performance from the augmented knowledge based on inter-relation between mRNA and miRNA** Dark blue bars represent the results from damaged graph and brown bar represents the one from augmented graph. Light blue bars indicate the AUCs of the original graph and reconstructed graph, respectively

Biological implication

Through the proposed model, the molecular signatures of miRNA and target genes, most associated with survival, were selected. First, miRNAs and gene features were separately selected from the prediction model based on intra-relation using independent data set, miRNA expression and gene expression, respectively. Then, miRNA and target gene pairs were selected from the prediction model based on inter-relation between miRNA and gene expression data. Figure 3-11 represents a heatmap of fold changes of selected miRNAs and genes, which are also belonging to selected miRNA-target gene pairs. The first column of Figure 3-11 shows the fold changes of gene expression from selected 11 genes and remaining columns represent the fold changes of

miRNA expression from selected 19 miRNAs. Blue cell in the figure indicates that gene expression or miRNA expression in the short-term survival group is under-expressed compared to the long-term survival group. Light blue cell in the heatmap represents non-target relation between miRNA and gene. Many of these miRNA and target gene pairs affect critical biological processes that are frequently dysregulated in cancer.

For instance, three miRNAs, hsa-mir-20a, hsa-mir-106a, and hsa-mir-221, were also identified as miRNA signatures that predicts survival in Glioblastoma (72). Hsa-mir-20a and hsa-mir-106a miRNAs were classified into the protective class and hsa-mir-221 was classified into the risk class in the previous study as well (72). The protective miRNAs were expressed at a higher level in the long-term survival group compared to the short-term survival group while the risky miRNAs were expressed at a higher level in the short-term group than in the long-term group. The risky and protective class of these miRNAs supports the fact that their functions being either promoting or inhibitory, respectively. Under-expression of hsa-mir-106a has been shown to be associated with poor patient survival in colon cancer and glioma (73-74). Target genes of hsa-mir-106a, BDH1, UPP1, TUSC2, and KMO, were over-expressed in the short-term survival group, which is a reverse pattern of expression in hsa-mir-106a. These genes play important roles that affect metabolic process, cell cycle, or nucleotide catabolic process in several cancers (14, 75-77). The miR-17~92 cluster, which contains hsa-mir-20a, was found to promote lung cancer growth in vitro, activated by c-myc and promote tumor angiogenesis (78). HFE, one of the selected target genes of

hsa-mir-20a, has been found to be associated with immune response in GBM and ovarian cancer (14, 79). Among selected miRNA and target gene pairs, other pairs were of interest because they could suggest some novel indirect mechanisms in GBM tumorigenesis.

Table 3-2 describes the selected gene features between short-term and long-term survival group. These gene lists were sorted by the *AUC_diff*, which calculated the difference between the original AUC with 11 gene features and the AUC without one gene among 11 gene features. The high value of *AUC_diff* means that the contribution of the gene feature, being excluded for calculating the *AUC_diff*, to the prediction model is high. RAGE showed the highest *AUC_diff*, 0.028, and *AUC_diff* of ATAD3A, 0.024, was secondly high among gene features (Table 3-2).

The RAGE pathway may play an important role in STAT3 induction in glioma-associated microglia and macrophages, a process that might be mediated through S100B (80). In addition, the under-expression of ATAD3A may be involved in the chemo-sensitivity of oligodendrogliomas and the transformation pathway (81).

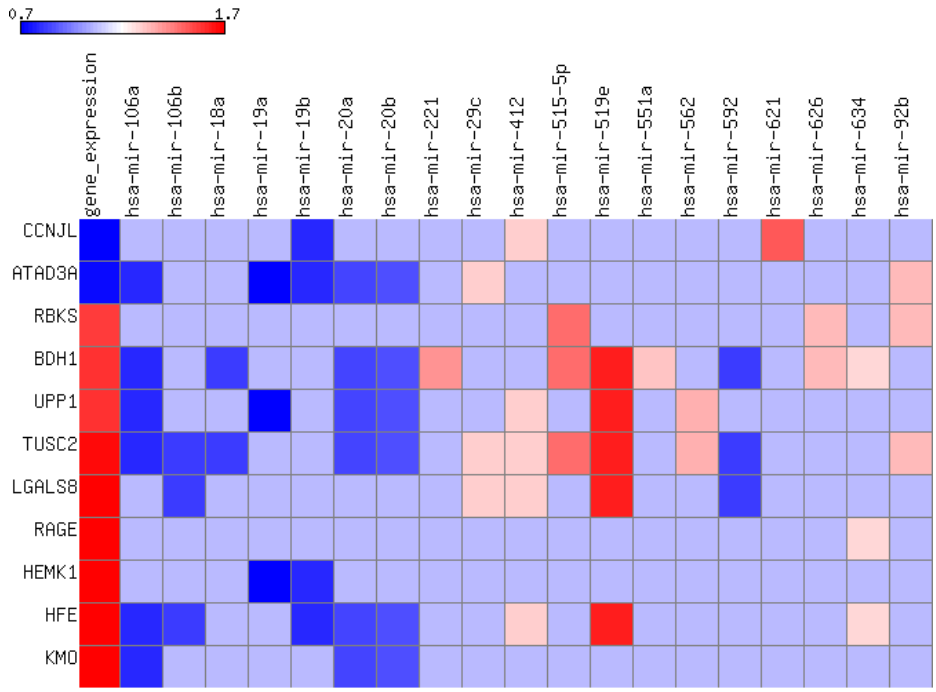


Figure 3−11. **Heatmap of selected miRNA and target gene pairs** The first column shows the fold changes of gene expression from selected 11 genes and remaining columns represent the fold changes of miRNA expression from selected 19 miRNAs. Blue cells indicate that gene expression or miRNA expression in the short-term survival group is under-expressed compared to the long-term survival group. Light blue cells represent non-target relation between miRNA and gene.

Table 3-2. Description of the selected gene features between short-term and long-term survival group in GBM

Gene	Region	Function	Up/down	AUC_diff
RAGE	14q32.31	Renal tumor antigen/threonine kinase activity/transferase activity	Up	0.028
ATAD3A	1p36.33	ATP binding/nucleotide binding	Down	0.024
HEMK1	3p21.31	DNA binding/N-methyltransferase activity	Up	0.012
KMO	1q43	Integral to membrane /kynurenine 3-monooxygenase activity	Up	0.012
RBKS	2p23.2	D-ribose metabolic process /ribokinase activity	Up	0.012
CCNJL	5q33.3	Nucleus/regulation of progression through cell cycle	Down	0.008
LGALS8	1q43	Extracellular space/sugar binding	Up	0.008
UPP1	7p12.3	Cytoplasm/nucleoside metabolic process/nucleotide catabolic process	Up	0.008
BDH1	3q29	3-hydroxybutyrate dehydrogenase activity/metabolic process/mitochondrial inner membrane/mitochondrial matrix	Up	0.004
HFE	6p22.1	Antigen processing and presentation/ immune response/ protein complex assembly	Up	0.004
TUSC2	3p21.31	Cell cycle/cell proliferation/cell-cell signaling/negative regulation of progression through cell cycle	Up	0.000

Comparison with other proposed methods for inter-relationship matrix

Despite the difficulty of developing an adequate measure to calculate the similarity matrix containing inter-relationship information between miRNA and gene expression, I implemented 4 measures, G_R_1, G_R_2, G_R_3, and G_R_4, and compared with the proposed method, G_R_5, in order to assess the benefit of the proposed one. G_R_1 was calculated by multiplication of correlation matrices from gene expression and miRNA expression. The method of G_R_2

was generated through the simple addition of two vectors, genes and miRNAs, for containing inter-relationship. On the other hand, the method of G_R_3 was calculated by removing miRNAs and genes, which were not belonging to the target relations, after simple addition of two vectors, genes and miRNAs. G_R_4 was focused on a targeted gene and considered multiple miRNAs targeting the specific gene when calculating the inter-relationship. In contrast to G_R_4, G_R_5, the proposed method in our study, was focused on a miRNA and considered multiple target genes from the specific miRNA.

Even though the performance of G_R_2 itself showed the best (AUC=0.828), the performance of G_A (AUC=0.868), integrating G_O (AUC=0.856) and G_R_5 (AUC=0.796), showed the best in our comparison scheme (Fig 3-12). This suggests that the method of G_R_5 has more partly complementary to the gene expression itself than the others so that it improves the prediction power when integrating with gene expression.

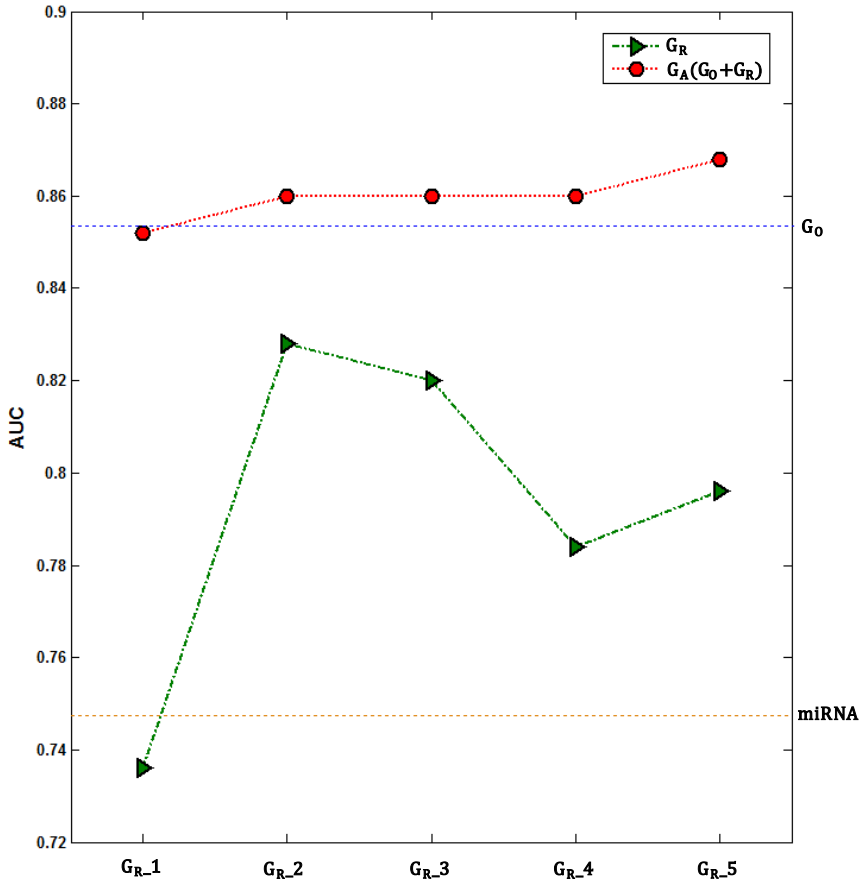


Figure 3−12. **Comparison of other proposed methods** Four measures, G_R_1, G_R_2, G_R_3, and G_R_4, were implemented and used for calculating G_A in order to assess the effect of the proposed method, G_R_5. The blue dotted line shows the AUC of original gene expression (G_O) and the orange dotted line represents the AUC of miRNA data alone.

DISCUSSION

In the chapter 3, the intra-relation of gene expression was reconstructed from inter-relation between miRNA and gene expression for prediction of short-term/long-term survival of GBM patients in order to provide a preliminary insight on the question that is how informative inter-relationship between miRNA and gene expression is when different levels of genomic dataset and valid genomic knowledge are available. Based on the results, the accuracy of our predictive model increases because of incorporation of information fused over genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression. New evidence suggests that genomic knowledge is complementary to the prediction power of explaining cancer phenotype even though biological experimental data such as gene expression has incomplete information. In addition, it is suggested that the utilization of external knowledge representing miRNA-mediated regulation of gene expression is substantially useful for elucidating the cancer phenotype since miRNAs regulate many genes associated with different biological processes such as development, stress response, apoptosis, proliferation, and tumorigenesis.

The present study underpins my on-going future work. It is expected that the next attempt will be more focused on how to utilize the information from 'intra-relation', the relation between different levels: from the genome level to epigenome, transcriptome, proteome, and further stretched to the phenome level. There might be other possible intra-relations between different layers of

genomic data such as ‘copy number alteration region - genes located in the alteration region,’ ‘DNA methylation site - specific genes regulated by promoter regions,’ *etc*. Therefore, when integrating multiple genomic data, it will be desirable that a framework will be capable of containing the inter-relationships between genomic features belonging to different layers of the biological system as genomic knowledge. Even though this study is limited to the prediction of short-term/long-term survival in GBM as a base task, the proposed framework can be applied to other cancer types or other clinical outcomes such as grade, stage, metastasis, *etc*. In addition, I could apply the proposed method to another layer of ‘intra-relation’ based on miRNA expression profiles together with ‘intra-relation’ between mRNAs.

CHAPTER 4

Knowledge bootstrapping: a graph-based integration with multi-omics data and genomic knowledge

INTRODUCTION

Previously, I have proposed an integrated framework that uses multi-layers of genomic data, copy number alteration, DNA methylation, gene expression, and miRNA expression, for the prediction of clinical outcomes in glioblastoma multiforme (GBM) and serous cystadenocarcinoma (82). The strengths of proposed approach was also highlighted as initiating its application using multiple scales (83). Nevertheless, in order to explain the phenotype of complex diseases, it is better way to incorporate the genomic knowledge when integrating multi-layers and heterogeneous genomic data.

DNA microarray technologies have been used to predict diagnosis or prognosis in several cancers as gene expression signatures, however, many studies within similar condition were shown to vary substantially between different studies (3, 84). Several methods with integrating genomic knowledge such as pathways or protein-protein interaction networks based on gene expression data have been developed to overcome variability of diagnostic or prognostic predictors and to increase their performances (85-89). These studies suggested that integrating gene expression data with genomic knowledge to construct pre-defined features results in higher performance in clinical outcome prediction and higher stability between different studies. Furthermore, incorporation of genomic knowledge offers obtained signatures from pre-defined geneset will be more interpretable and thus provides more insight into the complex molecular mechanisms in cancer.

However, none of previous studies provided the integrative framework for multi-omics data and genomic knowledge. Here, I propose a new integrative framework for multi-omics and genomic knowledge in order to better explain the phenotype of complex diseases (Fig. 4-1). As a pilot task, I used ovarian cancer data set from TCGA. Ovarian cancer (OV) is one of the most common gynecological malignancies, and is the 5th leading cause of cancer mortality in women in the United States (49). Understanding the molecular pathogenesis and underlying biology in cancer is expected to provide guidance for improved prognostic indicators and effective therapies.

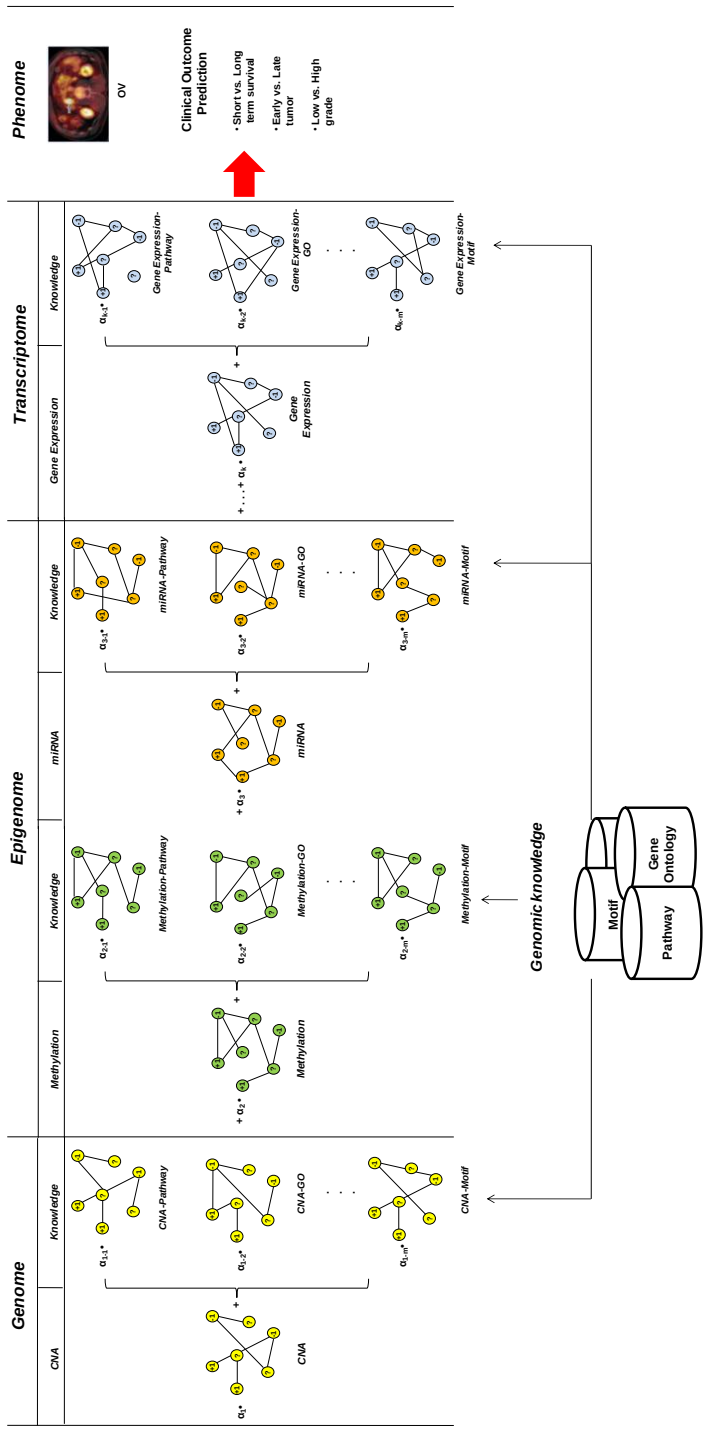


Figure 4−1. **Schematic overview of integration with multi-omics data and genomic knowledge**

MATERIALS AND METHODS

Data

Datasets were retrieved from the Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/) (Table 4-1). In order to direct map from feature to gene set, multi-omics data were summarized as a gene feature except for miRNA data. Copy number alteration (CNA) data were downloaded from level 3 in TCGA portal as segmentation results, and then divided into 23,228 gene features containing segmentation value when a gene was overlap in the segmented region. A matrix of DNA methylation was constructed by containing 9,129 genes, mapped to probes of DNA methylation chip. Gene expression data contains 12,043 genes.

A binary classification problem was set using the phenotype information from OV patients. Three sets of classifications are defined using ovarian cancer clinical outcomes, which are as follows: (1) *early stage (T1-T2) or late stage (T3-T4)*, (2) *low grade (G1-G2) or high grade (G3-G4)*, and (3) *short-term (< 3 y) or long-term (≥3 y) survival* (Table 4-2).

Table 4-1. Data description

Cancer type	Data type	Platform	# Features (*d*)
OV	CNA	Agilent SurePrint G3 Human CGH Microarray Kit 1x1M	23,228 genes
	Methylation	Infinium humanmethylation27 BeadChip	9,219 genes
	Gene expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043 genes
	miRNA	Agilent Human miRNA Microarray Rel12.0	799 miRNAs

Table 4-2. Clinical outcomes

Cancer type	Clinical outcome	# Samples (*n*)* (Neg/Pos)
OV	Short-term survival (survived less than three years) vs. long-term survival (survived longer than three years)	340 (147/193)
	Early stage (T1 or T2) vs. late stage (T3 or T4)	493 (39/454)
	Low grade (G1 or G2) vs. high grade (G3 or G4)	380 (19/361)

Genomic knowledge

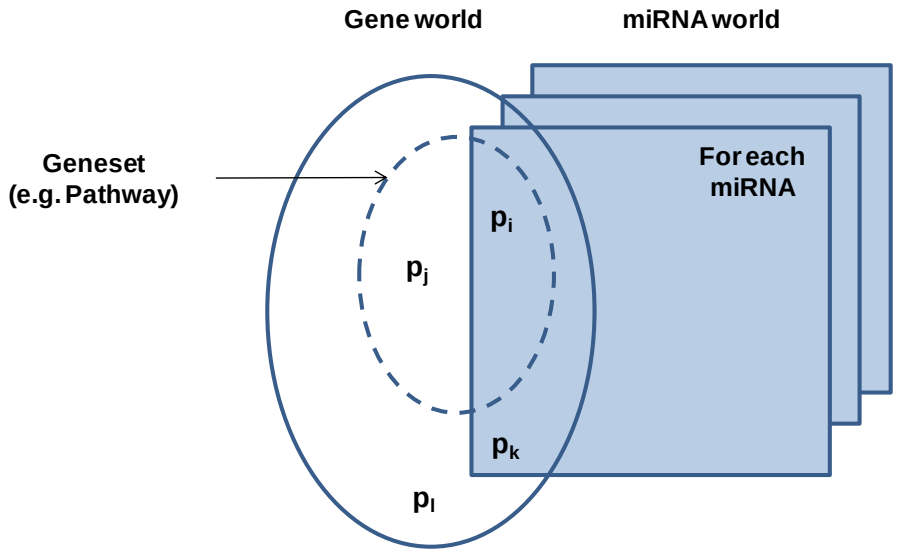
Pre-defined gene sets as genomic knowledge were downloaded from the Molecular Signatures Database (MsigDB 3.0) (Table 4-3)(90). Chromosomal position, pathway, motif, and Gene Ontology (GO) gene set were used for integrating with multi-omics data. Table 4-3 shows the description of genomic knowledge.

Table 4-3. Genomic Knowledge

Genomic Knowledge	# Gene sets	Source
Chromosomal positional gene set (C1)	326	MsigDB
Pathway (C2)	3,272	MsigDB
Motif (C4)	688	MsigDB
GO (C5)	1,454	MsigDB

Calculating gene sets for miRNA data

Each pre-defined gene set is a list of genes with relevant biological annotations. However, a miRNA itself has difficulties to directly map into gene sets since the member of gene set is a gene. Thus, new gene sets, which containing not genes but miRNAs, were needed to be defined in order to integrate with miRNA data and genomic knowledge relevantly. Suppose I want to test the enrichment of a gene set, which consists of genes, with respect to a specific miRNA, which consists of target genes. The numbers of genes in a gene set annotated (p_i) and not annotated (p_j) by the miRNA are used in the two by two contingency table along with the numbers of genes not in the gene set and are either annotated (p_k) or not annotated (p_l) with the miRNA, as shown in Figure 4-2. Enrichment is tested from the contingency table using a hypergeometric distribution. The p-value for the enrichment test from hypergeometric distribution of the random variable p is calculated from the cumulative probability of observing at least p_i our of $p_i + p_j$ times. Enrichment tests were conducted for each gene set, chromosomal positional gene set, pathway, motif, and GO gene set, respectively (p-value < 0.05).



	In Geneset	Not in Geneset
In miRNA	p_i	P_k
Not in miRNA	p_j	p_l

Figure 4−2. **Framework for calculating gene sets for miRNA data**

Clinical outcome classification

Graph-based semi-supervised learning I used a graph-based semi-supervised learning as a classification algorithm, which is a halfway learning scheme between supervised and unsupervised learning (41-44). Edges represent relations, more specifically similarities between samples that may be extracted from different genomic sources of CNA, methylation, gene expression, miRNA, etc. Different data sources produce different graphs. However, clinical outcome prediction can benefit by integrating diverse graphs from diverse genomic data sources or genomic knowledge, rather than relying only

on single sources that may have possible limitations, (i.e. incomplete information and noise). When data sources are presented as a graph form, combining multiple data sources can be done by employing a graph integration method (32-34).

Calculation similarity matrix containing genomic knowledge

The one of main problems in this study is to calculate the similarity matrix containing genomic knowledge. The underlying idea of the proposed measure is that genes do not act in isolation, and that complex diseases such as cancer are actually caused by the deregulation of complete processes or pathways. Thus, the feature unit of matrix of knowledge is gene set, i.e. pathway. The value of feature in matrix of knowledge was calculated by aggregation of values of the genes in the gene set (Fig. 4-3). After constructing a matrix of genomic knowledge, the similarity matrix was calculated using a Gaussian function and *k*-nearest-neighborhood method.

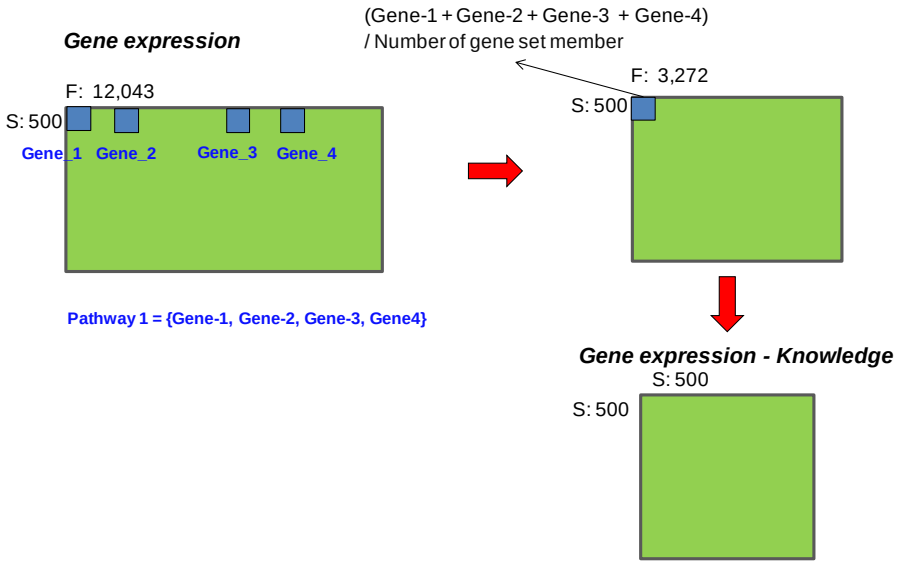


Figure 4-3. **Calculation similarity matrix containing genomic knowledge**

Integration of multiple graphs

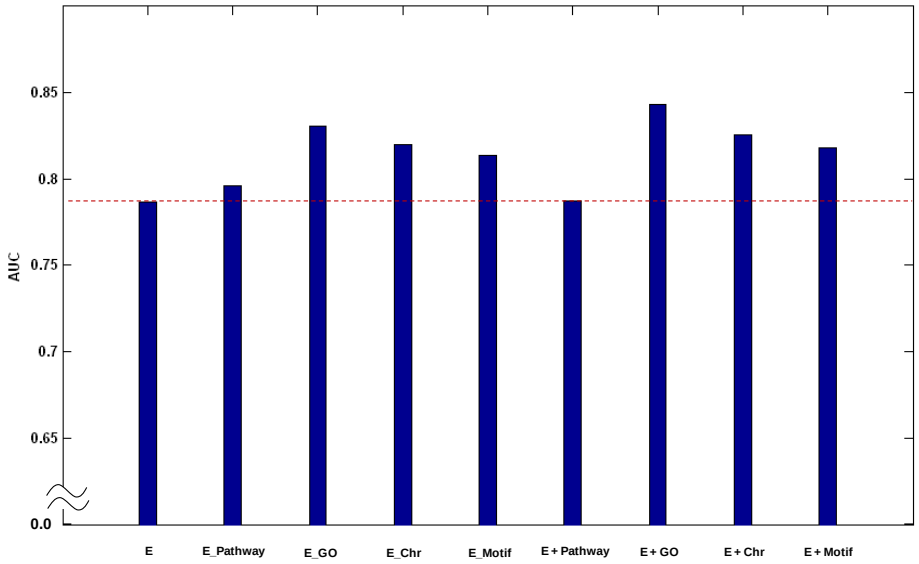
Integration of genomic data and genomic knowledge From genomic data and genomic knowledge, multiple graphs can be generated. Information from each graph is regarded as partially independent from and partly complementary to others. Therefore, it is not accurate enough to elucidate phenotype using only a single genomic data source belonging to a specific single layer. Reliability may be enhanced by integrating all available information sources using the method proposed by Tsuda *et al*. (2005), which has been re-validated on the previous study (82).

RESULTS

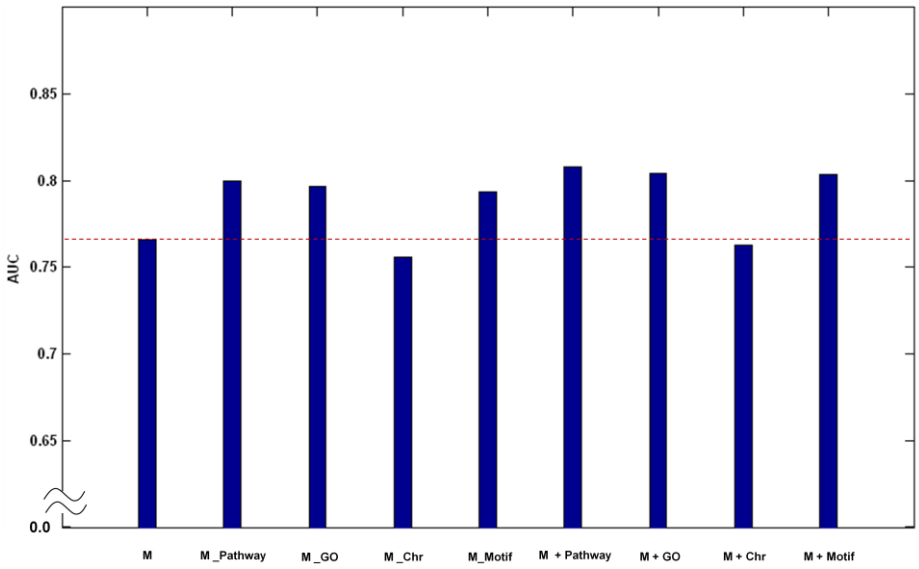
The receiver operating characteristic (ROC) curve plots sensitivity (true positive rate) as a function of 1-specificity (false positive rate) for a binary classifier system as its discrimination threshold is varied (46). An ROC score of 0.5 corresponds to random prediction, and an ROC score of 1.0 implies that the model succeeded in putting all of the positive examples before all of the negatives. For each problem, I calculated area under the curve (AUC) of ROC as a performance measure. Each experiment is repeated three times in order to estimate the variance of the measurement values and five-fold cross-validation was conducted.

Genomic knowledge integration effect

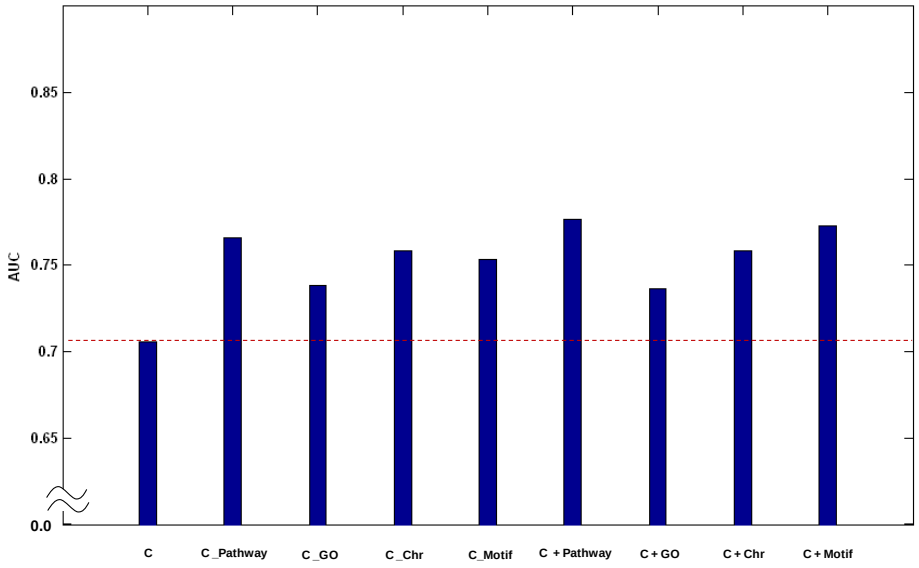
Figure 4-4, 4-5, 4-6 show the AUC performances on the three sets of classification problems of OV clinical outcomes. For the low vs. high grade classification, the SSL with gene expression data performed with an AUC of 0.7866. On the other hands, the integration with genomic knowledge generally performed better performances compared to the one with gene expression data only (Fig 4-4 (A)). Note that similar results were obtained for other genomic data, methylation, CNA, miRNA (Fig 4-4 (B, C, D)). For the other clinical outcomes, early vs. late stage and short-term vs. long-term survival, the results show a gradual increase in AUC by integration (Fig 4-5, 4-6). I found that the integration of various genomic knowledge increases the performance of clinical outcome prediction.



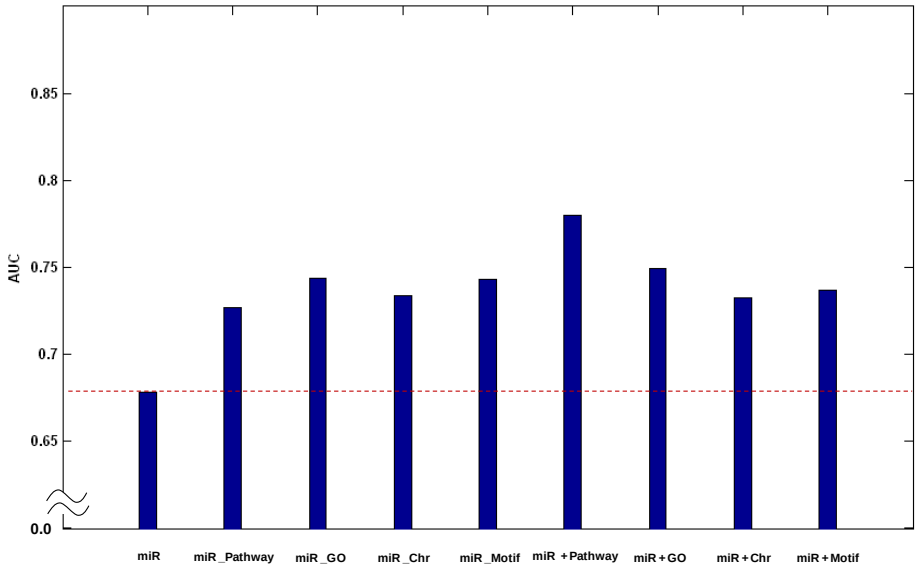
(A) Integration with gene expression and genomic knowledge



(B) Integration with methylation and genomic knowledge

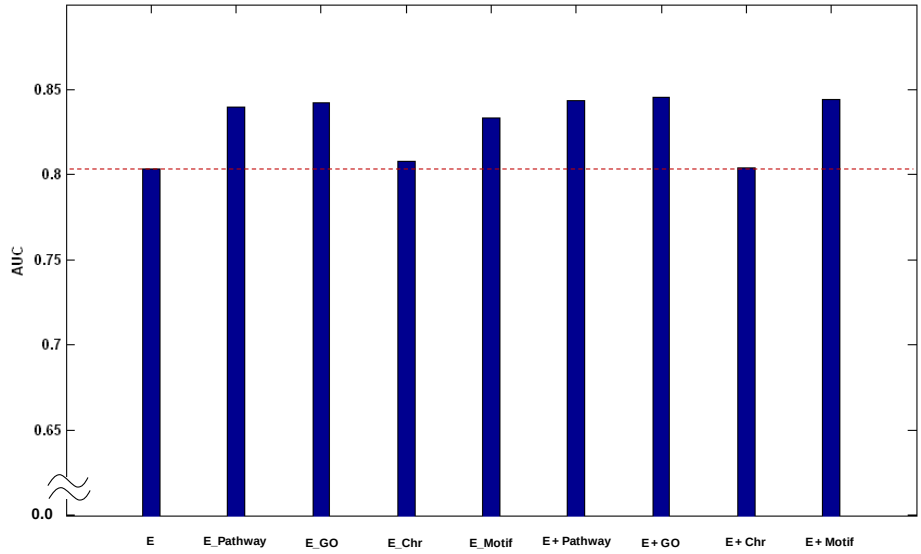


(C) Integration with CNA and genomic knowledge

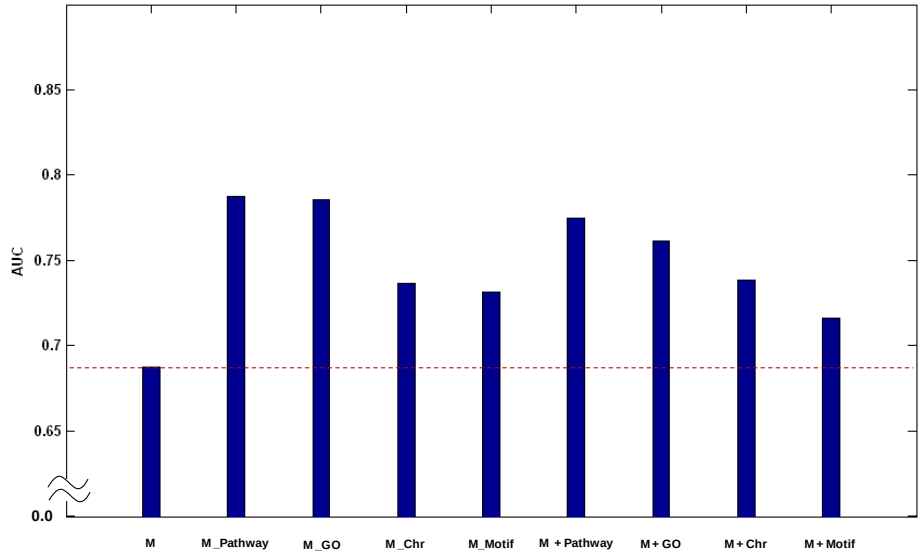


(D) Integration with miRNA and genomic knowledge

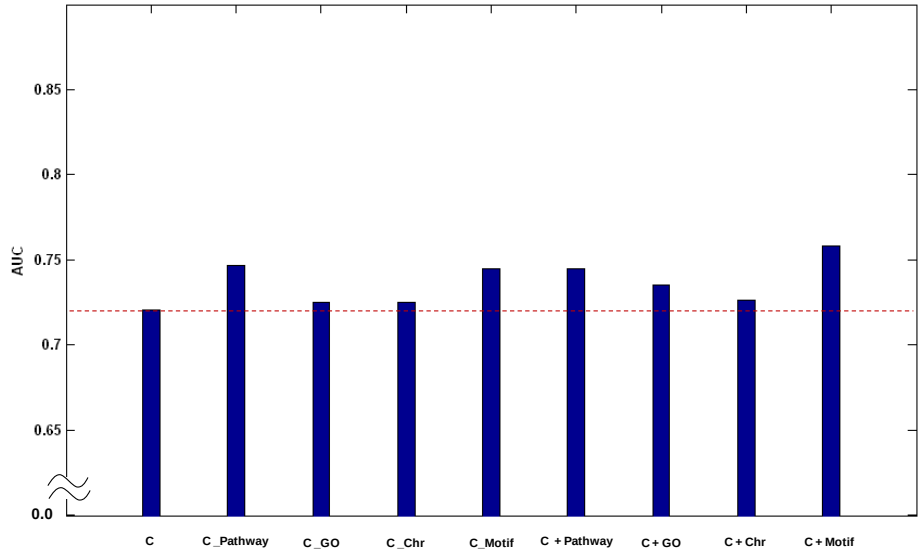
Figure 4−4. **Results of low vs. high grade outcome**



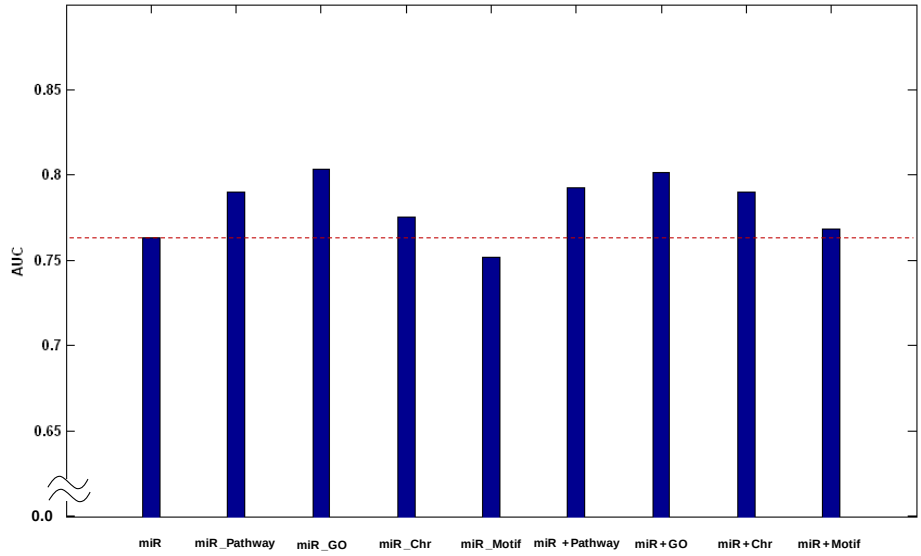
(A) Integration with gene expression and genomic knowledge



(B) Integration with methylation and genomic knowledge

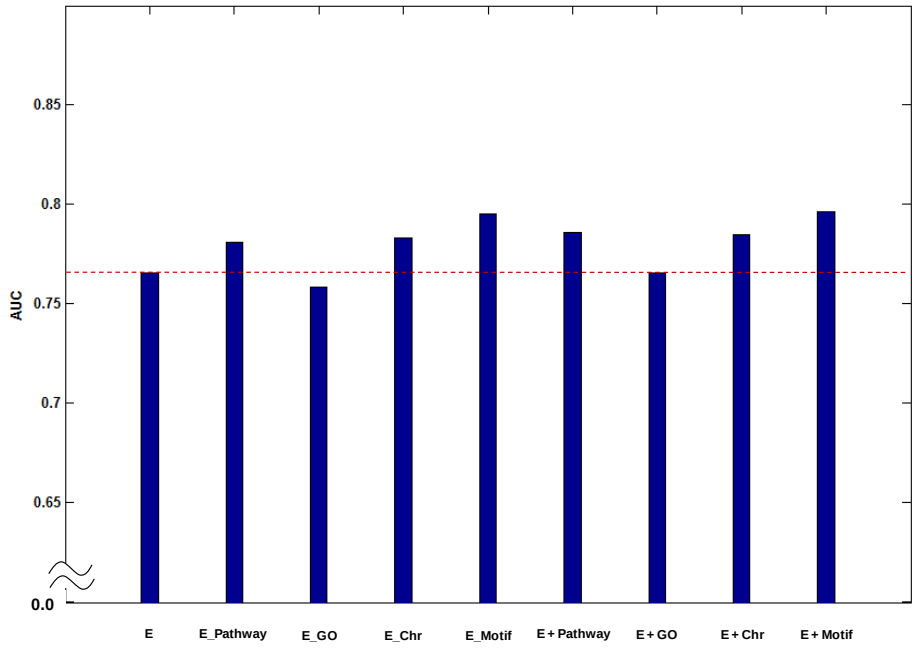


(C) Integration with CNA and genomic knowledge

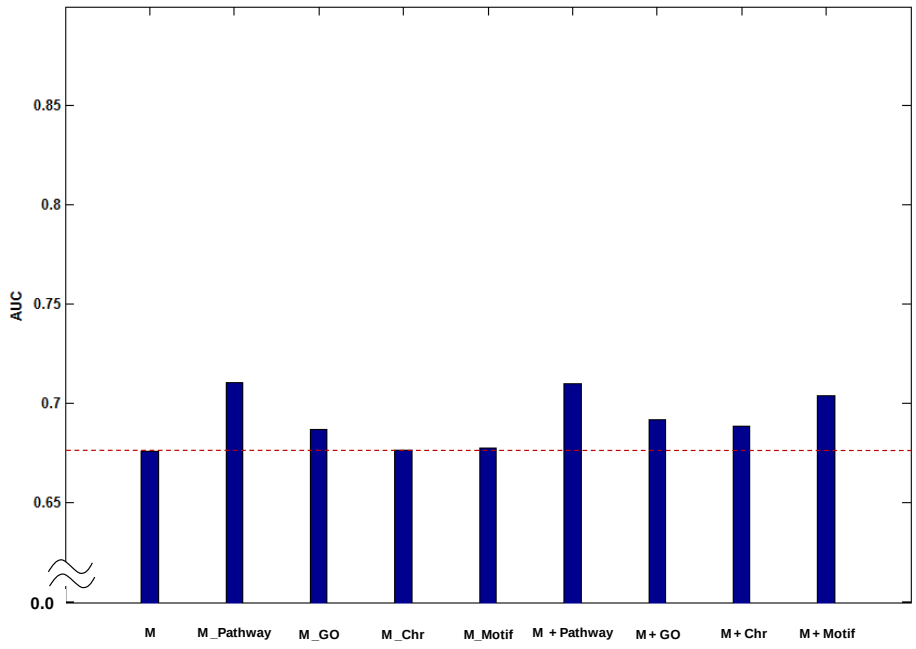


(D) Integration with miRNA and genomic knowledge

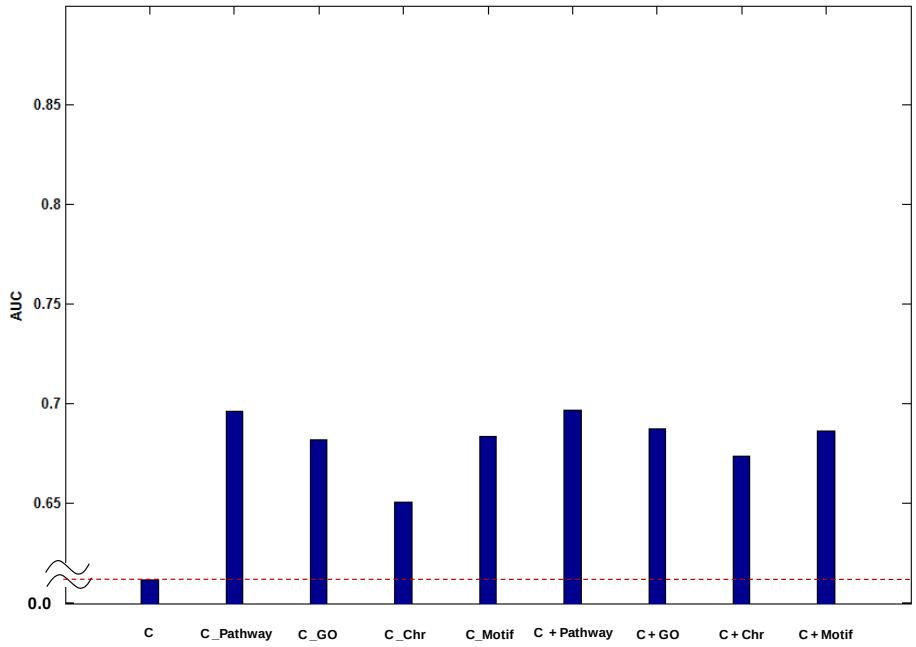
Figure 4-5. **Results of early vs. late stage outcome**



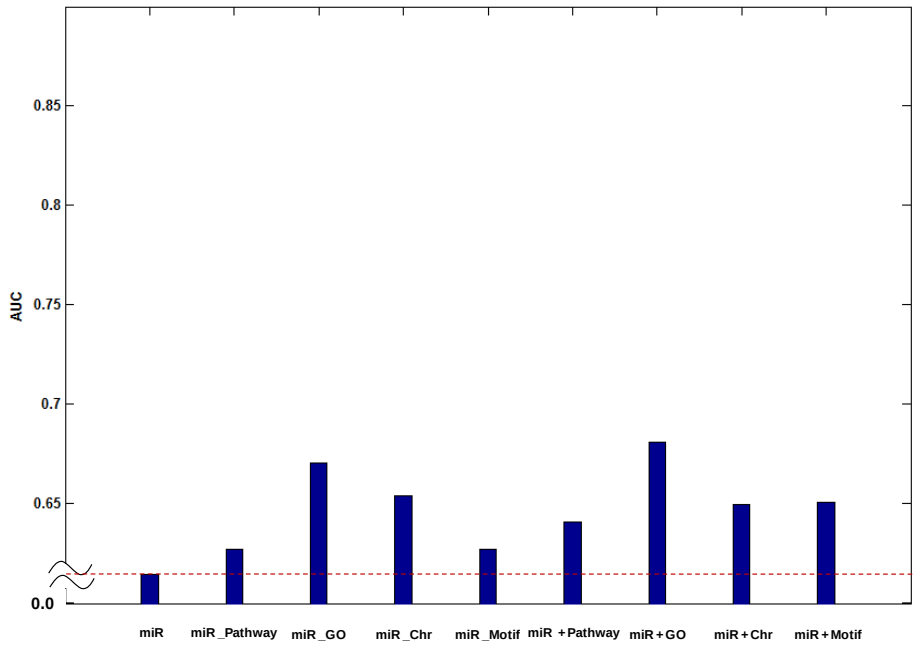
(A) Integration with gene expression and genomic knowledge



(B) Integration with methylation and genomic knowledge



(C) Integration with CNA and genomic knowledge

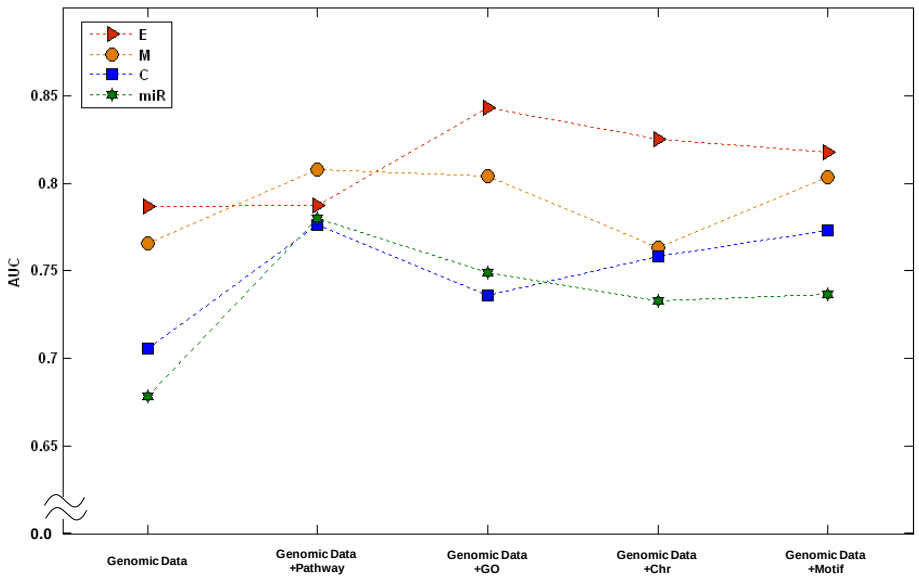


(D) Integration with miRNA and genomic knowledge

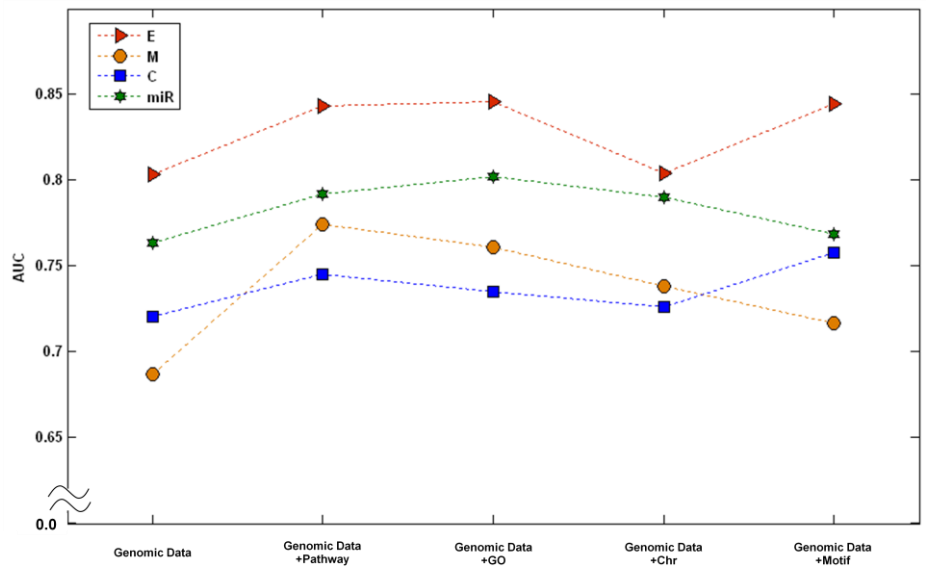
Figure 4−6. **Results of short-term vs. long-term survival outcome**

Relative contribution of genomic knowledge

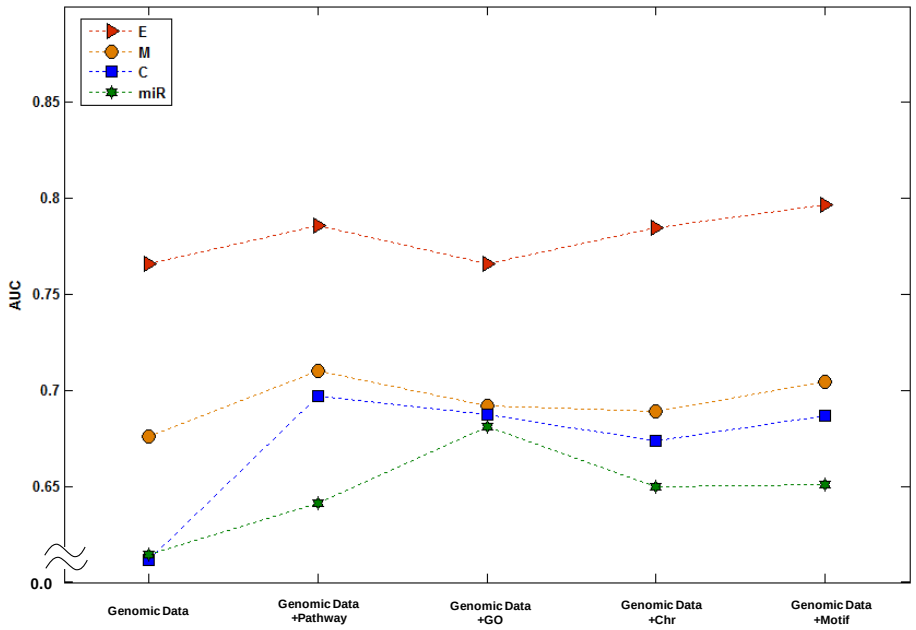
On the basis of the results of the computational experiments, some biological and clinical implications may be cautiously drawn. Figure 4-7 illustrates the following observations that show the level of contributions of genomic knowledge for 4 types of genomic data. For the low vs. high grade outcome, GO gene set performed best incorporating gene expression data. In contrast to gene expression data, CNA, methylation, and miRNA data showed that integration effect of pathway gene set was the best (Fig 4-7 (A)). For three cases of clinical outcome prediction, DNA methylation data consistently showed that pathway gene set was the best contributable in a model. However, the effect of genomic knowledge with other genomic data set, gene expression, CNA, and miRNA, was variable based on different clinical outcomes (Fig 4-7).



(A) Low vs. high grade



(B) Early vs. late stage



(C) Short-term vs. long-term survival

Figure 4−7. **Relative contribution of genomic knowledge**

DISCUSSION

In this chapter, I propose a new integrative framework for multi-omics and genomic knowledge in order to better explain the phenotype of complex diseases. In order to explain the phenotype of complex diseases, it is better way to incorporate the genomic knowledge when integrating multi-layers and heterogeneous genomic data. This study suggested that integrating gene expression data with genomic knowledge to construct pre-defined features results in higher performance in clinical outcome prediction and higher stability between different studies. Furthermore, incorporation of genomic knowledge offers obtained signatures from pre-defined geneset will be more interpretable and thus provides more insight into the complex molecular mechanisms in cancer.

To the best of my knowledge, none of previous studies provided the integrative framework for multi-omics data and genomic knowledge. A proposed framework has the advantage of scalability. Any kinds of genomic data or genomic knowledge can be integrated into the model as a graph. With integration of genomic knowledge, understanding the molecular pathogenesis and underlying biology in cancer is expected to provide better guidance for improved prognostic indicators and effective therapies.

GENERAL DISCUSSION

Since cancer is the phenotypic end-point of events cumulated through multiple levels of the biological system from genome to proteome, a single layer of biological information will not be sufficient to fully understand tumor behavior or the underlying biological mechanisms (12). Given multi-levels of data, information from a level to another may lead to some clues that help to uncover an unknown biological knowledge. Thus, integration of different levels of data can aid in extracting new knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of genomic data.

With abundance in genomic/clinical data, this study can be categorized into four chapters.

1. **Genomic data comparison: Which data is more informative?**

To wet-lab analysts, it concerns data generation that requires highly cost/time-demanding work and experienced facilities. To dry-lab analysts, it concerns selection of appropriate data source for more accurate prediction, avoiding unnecessary waste of computational resource. To provide a preliminary insight on the question, this study compares different types of genomic data using the state-of-the-art machine learning algorithm, Semi-Supervised Learning.

2. **Synergistic effect of different levels of genomic data for cancer clinical outcome prediction**

There have been many attempts in cancer clinical outcome prediction by suing a dataset from a number of molecular layers of biological system. Despite these efforts, however, it still remains difficult to elucidate the cancer phenotypes because the cancer genome is neither simple nor independent but rather complicated and dysregulated by multiple molecular mechanisms. Recently, heterogeneous types of genomic data, generated from all molecular levels of 'omic' dimensions from genome to phenome, for instance, *copy number variants* at the genome level, *DNA methylation* at the epigenome level, and *gene expression* and *microRNA* at the transcriptome level, have become available. In this study, I propose an integrated framework that uses multi-layers of heterogeneous genomic data for prediction of clinical outcomes in brain cancer (Glioblastoma multiforme, GBM) and ovarian cancer (Serous cystadenocarcinoma, OV).

3. Combining multi-layers of genomic data and inter-relationship between different layers of genomic features

The limitation of previous study is integration with multi-layers of genomic data for cancer clinical outcome prediction without considering of inter-relationship information between them. There are possible relationships between the sample features (attributes) belonging to different layers of genomic data such as 'miRNA-target genes,' 'copy number alteration region-genes located in the alteration region,' 'DNA methylation site-specific genes regulated by promoter regions,' etc. Therefore, when integrating multiple genomic data, it will be desirable that a framework will be capable of

containing the inter-relationships between sample features belonging to different layers of the biological system. This study can be categorized into three types of sub-studies.

4. Knowledge bootstrapping: a graph-based integration with multi-omics data and genomic knowledge

In order to explain the phenotype of complex diseases, it is better way to incorporate the genomic knowledge when integrating multi-layers and heterogeneous genomic data. Several methods with integrating genomic knowledge such as pathways or protein-protein interaction networks based on gene expression data have been developed to overcome variability of diagnostic or prognostic predictors and to increase their performances. However, none of previous studies provided the integrative framework for multi-omics data and genomic knowledge. Here, I propose a new integrative framework for multi-omics and genomic knowledge in order to better explain the phenotype of complex diseases.

Through proposed integrative framework, integration of different levels of data and genomic knowledge can aid in extracting new knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of genomic data.

Recently, TCGA started to generate the additional cancer genomic data for about 20 to 25 tumor types and planned finishing the data generation in the next few years as the second phase of the project. With abundance in multi-

layers of genomic and clinical data, our proposed integrative framework will be valuable for elucidating the underlying tumor behavior, eventually leading to more effective screening strategies and therapeutic targets in many types of cancer.

REFERENCES

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-7.

2. Fan X, Shi L, Fang H, Cheng Y, Perkins R, Tong W. DNA microarrays are predictive of cancer prognosis: a re-evaluation. Clin Cancer Res. 2010 Jan 15;16(2):629-36.

3. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530-6.

4. Roepman P, Wessels LF, Kettelarij N, Kemmeren P, Miles AJ, Lijnzaad P, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. Nat Genet. 2005 Feb;37(2):182-6.

5. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. Lancet. 2003 May 10;361(9369):1590-6.

6. Berchuck A, Iversen ES, Lancaster JM, Pittman J, Luo J, Lee P, et al. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. Clin Cancer Res. 2005 May 15;11(10):3686-96.

7. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005 Jun 9;435(7043):834-8.

8. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

9. Myllykangas S, Tikka J, Bohling T, Knuutila S, Hollmen J. Classification of human cancers based on DNA copy number amplification modeling. BMC Med Genomics. 2008;1:15.

10. Boeri M, Verri C, Conte D, Roz L, Modena P, Facchinetti F, et al. MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci U S A. 2011 Mar 1;108(9):3713-8.

11. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007 Nov 16;318(5853):1108-13.

12. Hanash S. Integrated global profiling of cancer. Nat Rev Cancer. 2004 Aug;4(8):638-44.

13. Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. Nature. 2008 Apr 3;452(7187):553-63.

14. TCGA. Integrated genomic analyses of ovarian carcinoma. Nature. 2011 Jun 30;474(7353):609-15.

15. Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 19;487(7407):330-7.

16. Network TCGA. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep 27;489(7417):519-25.

17. Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 4;490(7418):61-70.

18. TCGA Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-8.

19. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan 19;17(1):98-110.

20. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010 May 18;17(5):510-22.

21. Bolton KL, Chenevix-Trench G, Goh C, Sadetzki S, Ramus SJ, Karlan BY, et al. Association between BRCA1 and BRCA2 mutations and

survival in women with invasive epithelial ovarian cancer. JAMA. 2012 Jan 25;307(4):382-90.

22. Creighton CJ, Hernandez-Herrera A, Jacobsen A, Levine DA, Mankoo P, Schultz N, et al. Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma. PLoS One. 2012;7(3):e34546.

23. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012 Aug 24;337(6097):967-71.

24. Larman TC, DePalma SR, Hadjipanayis AG, Protopopov A, Zhang J, Gabriel SB, et al. Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A. 2012 Aug 28;109(35):14087-91.

25. Gravendeel LA, Kouwenhoven MC, Gevaert O, de Rooi JJ, Stubbs AP, Duijm JE, et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. Cancer Res. 2009 Dec 1;69(23):9065-72.

26. Li A, Walling J, Ahn S, Kotliarov Y, Su Q, Quezado M, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res. 2009 Mar 1;69(5):2091-9.

27. Qiu J, Noble WS. Predicting co-complexed protein pairs from heterogeneous data. PLoS Comput Biol. 2008 Apr;4(4):e1000054.

28. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics. 2004 Nov 1;20(16):2626-35.

29. Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. Bioinformatics. 2005 Jun;21 Suppl 1:i38-46.

30. Wu CC, Asgharzadeh S, Triche TJ, D'Argenio DZ. Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. Bioinformatics. 2010 Mar 15;26(6):807-13.

31. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003 Oct 17;302(5644):449-53.

32. Shin H, Lisewski AM, Lichtarge O. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics. 2007 Dec 1;23(23):3217-24.

33. Tsuda K, Shin H, Scholkopf B. Fast protein classification with multiple networks. Bioinformatics. 2005 Sep 1;21 Suppl 2:ii59-65.

34. Shin H, Tsuda K. Prediction of Protein Function from Networks. in Book: Semi-Supervised Learning, Edited by Olivier Chapelle, Bernhard Sch¨olkopf, Alexander Zien, MIT press. 2006;Chapter 20:339-52.

35. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003 Jun;34(2):166-76.

36. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998 Dec;9(12):3273-97.

37. Ohn JH, Kim J, Kim JH. Genomic characterization of perturbation sensitivity. Bioinformatics. 2007 Jul 1;23(13):i354-8.

38. Furnari FB, Fenton T, Bachoo RM, Mukasa A, Stommel JM, Stegh A, et al. Malignant astrocytic glioma: genetics, biology, and paths to treatment. Genes Dev. 2007 Nov 1;21(21):2683-710.

39. Salcman M, Kaplan R. Intracranial tumors in adults. In : Salcman M (ed) Neurology of brain tumors Williams & Wilkins, Baltimore. 1991:1339-52.

40. Saxena A, Robertson JT, Ali IU. Abnormalities of p16, p15 and CDK4 genes in recurrent malignant astrocytomas. Oncogene. 1996 Aug 1;13(3):661-4.

41. Chapelle O, Weston J, Scholkopf B. Cluster kernels for semi-supervised learning. Advances in Neural Information Processing Systems (NIPS). 2003;15(15):585-92.

42. Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In Proceedings of the Twenty-first International Conference on Machine Learning (ICML), Washington, DC, AAAI Press. 2003:912-9.

43. Belkin M. Regularization and Semi-supervised Learning on Large Graphs. In Proceedings of the 17th Annual Conference on Learning Theory (COLT) 3120 Lecture Notes in Computer Science. 2004:624-38.

44. Zhou D, Bousquet O, Weston J, Scholkopf B. Learning with local and global consistency. Advances in Neural Information Processing Systems (NIPS). 2004;16:321-8.

45. Chung FRK. Spectral Graph Theory. Number 92 in Regional Conference Series in Mathematics. 1997.

46. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem. 1996 Mar;20(1):25-33.

47. Demsar J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. 2006;7:1-30.

48. Albertson DG, Collins C, McCormick F, Gray JW. Chromosome aberrations in solid tumors. Nat Genet. 2003 Aug;34(4):369-76.

49. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin. 2009 Jul-Aug;59(4):225-49.

50. Marko NF, Toms SA, Barnett GH, Weil R. Genomic expression patterns distinguish long-term from short-term glioblastoma survivors: a preliminary feasibility study. Genomics. 2008 May;91(5):395-406.

51. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006 Jan 19;439(7074):353-7.

52. Shridhar V, Lee J, Pandita A, Iturria S, Avula R, Staub J, et al. Genetic analysis of early- versus late-stage ovarian tumors. Cancer Res. 2001 Aug 1;61(15):5895-904.

53. Waldman FM, DeVries S, Chew KL, Moore DH, 2nd, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. J Natl Cancer Inst. 2000 Feb 16;92(4):313-20.

54. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009 Jan 23;136(2):215-33.

55. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004 Jan 23;116(2):281-97.

56. van Rooij E, Sutherland LB, Liu N, Williams AH, McAnally J, Gerard RD, et al. A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure. Proc Natl Acad Sci U S A. 2006 Nov 28;103(48):18255-60.

57. Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. Science. 2004 Jan 2;303(5654):83-6.

58. Raver-Shapira N, Marciano E, Meiri E, Spector Y, Rosenfeld N, Moskovits N, et al. Transcriptional activation of miR-34a contributes to p53-mediated apoptosis. Mol Cell. 2007 Jun 8;26(5):731-43.

59. Marsit CJ, Eddy K, Kelsey KT. MicroRNA responses to cellular stress. Cancer Res. 2006 Nov 15;66(22):10843-8.

60. Schmittgen TD. Regulation of microRNA processing in development, differentiation and cancer. J Cell Mol Med. 2008 Oct;12(5B):1811-9.

61. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009 Jan;37(Database issue):D105-10.

62. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12963-8.

63. McCabe MT, Brandes JC, Vertino PM. Cancer DNA methylation: molecular mechanisms and clinical implications. Clin Cancer Res. 2009 Jun 15;15(12):3927-37.

64. Schuebel KE, Chen W, Cope L, Glockner SC, Suzuki H, Yi JM, et al. Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. PLoS Genet. 2007 Sep;3(9):1709-23.

65. Miles GD, Seiler M, Rodriguez L, Rajagopal G, Bhanot G. Identifying microRNA/mRNA dysregulations in ovarian cancer. BMC Res Notes. 2012;5:164.

66. Liu H, Brannon AR, Reddy AR, Alexe G, Seiler MW, Arreola A, et al. Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma. BMC Syst Biol. 2010;4:51.

67. Jayaswal V, Lutherborrow M, Yang YH. Measures of Association for Identifying MicroRNA-mRNA Pairs of Biological Interest. Plos One. 2012 Jan 11;7(1).

68. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005 Jan 14;120(1):15-20.

69. Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT. miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics. 2006;7:411.

70. Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak. 2006;6:27.

71. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007 Oct 1;23(19):2507-17.

72. Srinivasan S, Patric IR, Somasundaram K. A ten-microRNA expression signature predicts survival in glioblastoma. PLoS One. 2011;6(3):e17438.

73. Diaz R, Silva J, Garcia JM, Lorenzo Y, Garcia V, Pena C, et al. Deregulated expression of miR-106a predicts survival in human colon cancer patients. Genes Chromosomes Cancer. 2008 Sep;47(9):794-802.

74. Zhi F, Chen X, Wang SN, Xia XW, Shi YM, Guan W, et al. The use of hsa-miR-21, hsa-miR-181b and hsa-miR-106a as prognostic indicators of astrocytoma. European Journal of Cancer. 2010 Jun;46(9):1640-9.

75. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012 Apr 4.

76. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC, et al. The genetic landscape of the childhood cancer medulloblastoma. Science. 2011 Jan 28;331(6016):435-9.

77. Durinck S, Ho C, Wang NJ, Liao W, Jakkula LR, Collisson EA, et al. Temporal Dissection of Tumorigenesis in Primary Cancers. Cancer Discov. 2011 Jul;1(2):137-43.

78. Bonauer A., S. D. The microRNA-17-92 cluster: still a miRacle? Cell Cycle. 2009;8:3866–73.

79. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008 Sep 26;321(5897):1807-12.

80. Zhang L, Liu W, Alizadeh D, Zhao D, Farrukh O, Lin J, et al. S100B attenuates microglia activation in gliomas: possible role of STAT3 pathway. Glia. 2011 Mar;59(3):486-98.

81. Hubstenberger A, Labourdette G, Baudier J, Rousseau D. ATAD 3A and ATAD 3B are distal 1p-located genes differentially expressed in human glioma cell lines and present in vitro anti-oncogenic and chemoresistant properties. Experimental Cell Research. 2008 Sep 10;314(15):2870-83.

82. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. Journal of Biomedical Informatics. 2012;Accepted for publication.

83. Lussier YA, Li H. Breakthroughs in genomics data integration for predicting clinical outcome. J Biomed Inform. 2012 Dec;45(6):1199-201.

84. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005 Feb 19-25;365(9460):671-9.

85. Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinformatics. 2010;11:277.

86. Ma S, Shi M, Li Y, Yi D, Shia BC. Incorporating gene co-expression network in identification of cancer prognosis markers. BMC Bioinformatics. 2010;11:271.

87. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009 Feb;27(2):199-204.

88. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008 Nov;4(11):e1000217.

89. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

90. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;27(12):1739-40.

국문 초록

서론: 의학의 발전과 더불어 암의 진단과 치료에 있어서 많은 발전이 이루어 졌지만, 아직 정확한 암의 조기 진단 및 예후 파악이 쉽지 않아 암 환자들이 적절한 시기에 치료를 받지 못하고 죽음에 이르게 되는 경우가 많다. 최근에는 암의 진단과 치료를 위해 마이크로어레이 기술을 이용하여 암에 의해 변화하는 대량의 유전자 발현 패턴을 조사하고, 이렇게 얻어진 자료에 전산, 통계학적 기법을 적용하여 진단과 임상 예후 예측에 응용하려는 시도들이 많이 있다. 하지만, 암은 다양한 생물학적인 메커니즘들을 통해서 발생 및 발달하기 때문에, 보다 엄밀히 진단 및 임상 예후 예측을 하기 위해서는 다양한 레벨의 유전체 데이터로부터 통합 분석의 필요성이 증대되고 있다.

방법: 미국에서 암의 복잡한 메커니즘을 규명하기 위해서 시작 된 대규모 프로젝트 (the Cancer Genome Atlas, TCGA)로부터 다양한 암 환자들에 대해서 다계층 이종 유전체 데이터 및 임상데이터들이 공개되기 시작했다. 본 연구에서는 기계학습을 이용해서 다양한 분자 수준의 이종 유전체 데이터들을 통합해서 암의 진단 및 임상 예후 예측을 위한 분류 모델을 제시 한다. 그래프 기반의 통합 모델을 제시 함으로써, 이종 유전체 데이터들은 각각 그래프 형태로 변환 된다. 다계층 유전체 데이터로부터 구성된 그래프들은 최적화

방법론을 이용해서 최적의 그래프로 통합되어 암의 진단 및 임상 예후 예측에 사용된다.

결과: 제안하는 통합 분석 모델의 타당성을 검증하기 위해서, 다형 교아종과 혈청 낭종암 환자들로부터 copy number, gene expression, methylation, miRNA 데이터를 통합하여 환자들의 진단 및 임상 예후 예측을 수행하였다. 단일 유전체 데이터로 예측을 할 때 보다 다계층 유전체 데이터를 통합하였을 때 예측률이 더 높았다. 또한, 쉽게 확장 가능한 그래프 기반의 통합 모델을 제시함으로써, 다양한 유전체 지식도 함께 모델에 쉽게 통합 될 수 있다. 마찬가지로, 다계층 유전체 데이터에 잘 알려진 여러 유전체 지식을 같이 통합 하였을 때 암 환자들의 진단 및 임상 예후 예측이 더 잘 되는 것을 확인하였다.

결론: 암의 진단과 치료를 위해 다계층 유전체 데이터를 이용하여 암에 의해 변화하는 대량의 분자들의 발현 패턴을 연구함으로써, 보다 정확한 진단과 예후 예측이 가능한 것을 본 연구를 통해서 확인하였다. 결론적으로 다계층 유전체 데이터를 통합해서 분석함에 따라 암의 복잡한 메커니즘의 이해가 증가 될 뿐만 아니라, 더 나아가서 암과 연관 된 다계층 분자 후보 군을 제시함으로써, 암 치료제 개발에 크게 기여할 수 있을 것이다.

주요어 : 통합분석, 다계층 유전체 데이터, 임상예후 예측, 다형교아종, 혈청 낭종암
학 번 : 2006-22113