2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

Multi-Block and Multi-Task Learning for Integrative Genomic Study

Mingon Kang [1*], Dong-Chul Kim [2], Chunyu Liu[3], Baoju Zhang[4], Xiaoyong Wu[4], and Jean Gao [1*]
[1]*Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA 76019*
[2]*Department of Computer Science, University of Texas - Pan American, Edinburg, TX, USA 78539*
[3]*Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA 66012*
[4]*College of Electronics and Communication Engineering, Tianjin Normal University, Tianjin, China 300074*
[*]*Correspondence: mkang@uta.edu, gao@uta.edu*

***Abstract*—The importance of an integrative genomic study is steadily increasing in an emerging era of various high-throughput genomic data. Mechanisms of human diseases consist of complex interactions of multiple biological processes such as genetic, epigenetic, and transcriptional regulation. The collection of the multiple genomic data that represents the multiple processes is called 'multi-block data'. The multi-block data profiled from human disease samples provide comprehensive global snapshots of the diseases. Due to the rapid development of high-throughput technologies, the integrative genomic study using the multi-block data has been more highlighted than ever. However, in spite of its importance, there are only a few methodologies that can analyze such data. In this paper, we propose a novel Multi-Block and Multi-Task Learning (MBMTL) method for the integrative genomic study. We consider Single Nucleotide Polymorphism (SNP), Copy Number Variation (CNV), DNA methylation, and gene expression data as the multi-block data from four group samples of three major psychiatric disorders as well as data from a normal control. MBMTL identifies biomarkers that play important roles in explaining mechanisms of the human diseases from the multi-block data. We also take a multi-task problem into account so that we can identify different functions of the mechanisms. The performance of the proposed MBMTL was assessed by comparing it to a number of existing multi-block methods through simulation studies. We applied MBMTL to the multi-block data of the major psychiatric disorder samples.**

I. Introduction

An integrative genomic study has been in the spotlight more than ever due to the remarkable advancements of the high-throughput technologies. The efficient acquisition of multiple genomic data by the high-throughput technologies provides global snapshots of multiple biological processes such as genetic, epigenetic, and transcriptional regulation. The genomic data include Single Nucleotide Polymorphism (**SNP**), Copy Number Variation (**CNV**), **DNA** Methylation, and gene expression.

Mechanism of human diseases consist of complex biological processes that involve multiple interplays among genetic and biochemical components. Therefore, demands for novel integrative approaches, which can perform integrated analysis with the multiple genomic data, are increasingly growing. Also recent striking achievement in improvements of accuracy and resolution of the genomic data naturally bring out novel integrative approaches.

Nevertheless, most research has focused solely on examination using a single type of genomic data yet. Genetic variations that are associated to human diseases or gene expressions has been identified in Genome-Wide Association Studies (**GWAS**) and expression Quantitative Trait Loci (**eQTL**) mapping studies using **SNP** [1], [2]. In addition to the genetic variations, Copy Number Variations (**CNV**s), which are alterations of the DNA regions, have been studied [3], [4]. DNA methylation, which is an epigenetic modification, also has been highlighted as a causal factor that controls gene expression levels [5], [6].

An integrative genomic model that take the multiple genomic data into account has been widely attempted in recent research [7], [8]. The integrative genomic model is typically comprised of the relationships of **SNP**, **CNV**, **DNA** methylation, and gene expression data. The integrative genomic model is then analyzed by statistical approaches. In the model, the lack of the genetic interpretation by the variations of **SNP** and **CNV** can be explained by the epigenetic modification of **DNA** methylation when expressing the changes of gene expressions.

For the multi-block analysis of the multiple genomic data, Partial Least Squares (**PLS**) and Generalized Canonical Correlation Analysis (**GCCA**) based approaches have been proposed [9]–[11]. A sparse method of **PLS** by penalizing both features and sample dimensions identifies '*regulatory modules*' in the integrative genomic model [9]. Sparse Generalized Canonical Correlation Analysis (**SGCCA**) was proposed to analyze a hybrid model that combines **GWAS** and **eQTL** [10]. The data integration approach by utilizing multiple feature selection methods such as Principal Component Analysis (**PCA**), **PLS**, and **LASSO** was also proposed [11]. However, the approaches do not account for the different models between human diseases and a control, which is very important when identifying the biomarkers of diseases.

To tackle this problem, we propose a novel multi-block and multi-task learning (MBMTL) method. The proposed method **MBMTL** is designed to fulfill the following requirements:

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.21

- the method needs to deal with the multiple genomic data for the integrative genomic model.
- the method needs to identify the different functions of diseases and a control.

First, **MBMTL** is capable of the comprehensive analysis using the multiple genomic data in the integrative genomic model. We consider the typical integrative genomic model which includes **SNP**, **CNV**, **DNA** methylation, and gene expression. Secondly, **MBMTL** handles the multiple models of different diseases as well as a control. In this paper, we analyzed three major psychiatric disorder and control samples, and identified the various different functions.

II. Methods

A. Notation

We assume that there are L task models, where multi-block data of an identical structure exist but represent different models. In the l-th task, there are multiblock data that include J numbers of blocks, $\mathbf{X}^l = \{\mathbf{X}_1^l, ..., \mathbf{X}_J^l\}, 1 \leq l \leq L$. A block data $\mathbf{X}_j^l$ is a P_j-dimensional mean-zero matrix obtained from N_l samples, $\mathbf{X}_j^l \in \Re^{N_l \times P_j}, N = \sum_{l=1}^{L} N_l$. In this paper, we consider four block data of **SNP**, **CNV**, **DNA** methylation, and gene expression, which are denoted by $\mathbf{X}_1^l$, $\mathbf{X}_2^l$, $\mathbf{X}_3^l$, and $\mathbf{X}_4^l$, respectively ($\mathbf{X}^l = \{\mathbf{X}_1^l, \mathbf{X}_2^l, \mathbf{X}_3^l, \mathbf{X}_4^l\}$). We use i for an index of samples, and j and k for an index of multi-block through this paper.

B. Multi-Block and Multi-Task Learning

Multi-Block and Multi-Task Learning (**MBMTL**) optimizes linear projections of the multi-block data that maximize the total squared correlation between them. In this multi-block method, there are two relations to be considered – outer and inner relation. The outer relation is a linear projection which is a principal component of the block data. The linear projection $\mathbf{v}_j^l$ of the j-th block and the l-th task is defined:

$$\mathbf{v}_j^l = \mathbf{X}_j^l \mathbf{a}_j^l, \tag{1}$$

where $\mathbf{a}_j^l$ is a loading vector ($\mathbf{a}_j^l \in \Re^{P_j}$). The linear projection is called '*latent variable*'.

The inner relation is a connectivity between the latent variables. The inner relation should be given by a expert. A matrix for the inner relation $\mathbf{D} = \{d_{jk} | d_{jk} \in \{0, 1\}, 1 \leq j, k \leq J\}$ is defined as $d_{jk} = 1$ if the block j and k are connected, or otherwise 0. In the inner relation, a latent variable $\mathbf{v}_j^l$ is a summary vector of the latent variables which is connected to the j-th block:

$$\mathbf{v}_j^l \approx \sum_{k=1}^{J} d_{jk} b_k^l \mathbf{X}_k^l \mathbf{a}_k^l. \tag{2}$$

For instance, the matrix $\mathbf{D}$ for the proposed integrative genomic model is defined as (see Fig. 1):

$$D = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}. \tag{3}$$

The latent variable of gene expression block $\mathbf{v}_4^l$ is expressed by a linear combination of **SNP**, **CNV**, and **DNA** methylation:

$$\mathbf{v}_4^l \approx b_1^l \mathbf{X}_1^l \mathbf{a}_1^l + b_2^l \mathbf{X}_2^l \mathbf{a}_2^l + b_3^l \mathbf{X}_3^l \mathbf{a}_3^l. \tag{4}$$

MBMTL also considers a multi-task problem in the multi-block analysis. The multi-task problem considers several related tasks, but shares common features. For example, we can assume that psychiatric diseases such as bipolar disorder, major depression, and schizophrenia mainly share the mechanism that causes the diseases. However, the common mechanism may have different functions depending on the disease. In other words, a sparse subset of a loading vector $\mathbf{a}_j^l$ represents the common features across the related tasks, where the coefficient values of the loading vector may be different depending on the task. For the multi-task setting, we defines a loading matrix $\mathbf{A}_j$, which is a $P_j \times L$ dimensional matrix, $\mathbf{A}_j = \{\mathbf{a}_j^1, ..., \mathbf{a}_j^L\}$.

The objective function of **MBMTL** is defined as:

$$\begin{aligned} \text{argmax}_{\mathbf{A}_{j,1 \leq j \leq J}} \quad & \sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{k=1, j \neq k}^{J} d_{jk} \, \text{corr}(\mathbf{X}_j^l \mathbf{a}_j^l, \mathbf{X}_k^l \mathbf{a}_k^l)^2 \\ \text{s. t.} \quad & \mathbf{a}_j^{l\top} \mathbf{X}_j^{l\top} \mathbf{X}_j^l \mathbf{a}_j^l = 1, \\ & ||\mathbf{A}_j||_{2,1}^2 < t, \\ & 1 \leq l \leq L, \quad 1 \leq j \leq J, \quad 0 \leq t \end{aligned} \tag{5}$$

where $||\mathbf{A}_j||_{2,1}$ is $\ell_{2,1}$-norm of $\mathbf{A}_j$, i.e., $||\mathbf{A}_j||_{2,1} = \sum_{m=1}^{L} \sqrt{\sum_{n=1}^{P_j} \left(A_j^{[mn]}\right)^2}$. The $\ell_{2,1}$-norm regularization term computes the ℓ_2-norm of $\mathbf{A}_j^{[*n]}$ across the features and then ℓ_1-norm of the vector. $\mathbf{A}_j^{[m*]} \in \Re^L$ and $\mathbf{A}_j^{[*n]} \in \Re^{P_j}$ represent the m-th row and the n-th column of $\mathbf{A}_j$ respectively. Computing both the ℓ_2 and ℓ_1-norm enables us to select common features across the tasks (see Fig. 2).

The objective function can be converted to an alternating least squares problem [12]. The optimal solution is estimated by performing the following alternative equations iteratively on each block:

$$\mathbf{A}_j = \text{argmin} \sum_{l=1}^{L} \sum_{j=1}^{J} \left(\mathbf{h}_j^l - \mathbf{X}_j^l \mathbf{a}_j^l\right)^2 + \lambda_j \parallel \mathbf{A}_j \parallel_{2,1}^2,$$

where $\mathbf{h}_j^l = \sum_{k=1}^{J} d_{jk} \, \text{corr}(\mathbf{X}_j^l \mathbf{a}_j^l, \mathbf{X}_k^l \mathbf{a}_k^l) \mathbf{X}_k^l \mathbf{a}_k^l$. On the iteration of the j-th block, the alternative least squares problem optimizes $\mathbf{A}_j$ given $\mathbf{h}_j^l$. This problem, which is equivalent to the Joint Feature Selection problem of Argyriou et al, can

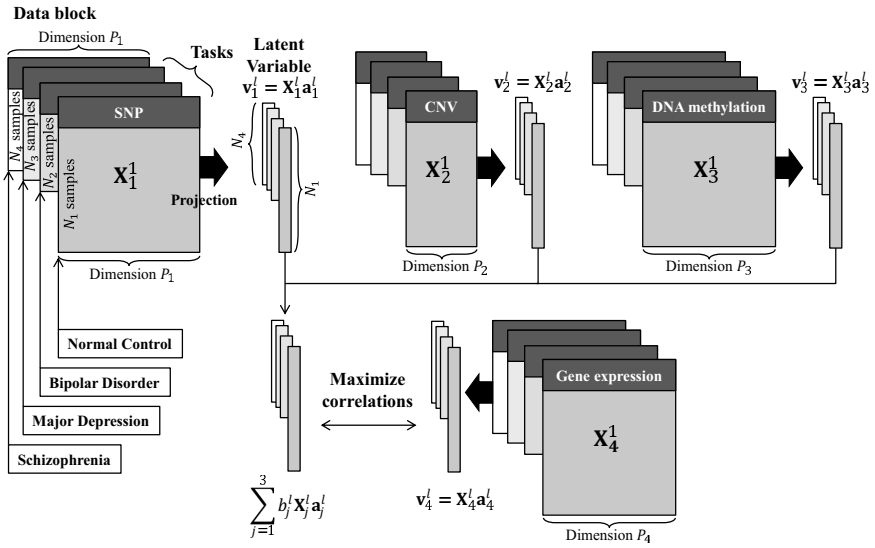


Figure 1: The conceptual graphical representation of the integrative genomic model in **MBMTL**.

Algorithm 1 Multi-Block and Multi-Task Learning

Normalize $\mathbf{X}_j^l$
Initialize $\mathbf{a}_j^l = \mathbf{1}/N_j$
repeat
  for $j := 1$ **to** J **do**
    for $k := 1$ **to** J **do**
      for $l := 1$ **to** L **do**
        $\mathbf{v}_j^l = \sum_{k=1}^{J} d_{jk} \operatorname{corr}\left(\mathbf{X}_j^l \mathbf{a}_j^l, \mathbf{X}_k^l \mathbf{a}_k^l\right) \mathbf{X}_k^l \mathbf{a}_k^l$
      end for
      Compute $\mathbf{A}_j$ by Eq. (6)
      Normalize $\mathbf{a}_j^l$ by Eq. (6)
    end for
  end for
until $\sum_{j=1}^{J} \| \mathbf{A}_j \|_{2,1}^2$ converges

be efficiently solved [13]. Then, $\mathbf{a}_j^l$ is normalized in order to satisfy the first constraint of (5):

$$\mathbf{a}_j^l = \sqrt{N_l}\mathbf{a}_j^l / |\mathbf{X}_j^l \mathbf{a}_j^l|. \tag{6}$$

The algorithm is described in Algorithm 1.

III. Simulation Study

The performance of the proposed method **MBMTL** was assessed by a number of simulated data of various complexities. In the simulation study, we compared the performance with other existing multi-block methods such as Sparse Canonical Correlation Analysis (**SCCA**), and Sparse Generalized Canonical Correlation Analysis (**SGCCA**). Then, we performed **MBMTL** to human brain data of psychiatric disorders in order to unveil the complex mechanisms using the genomic multi-block data and compared it to a health control.

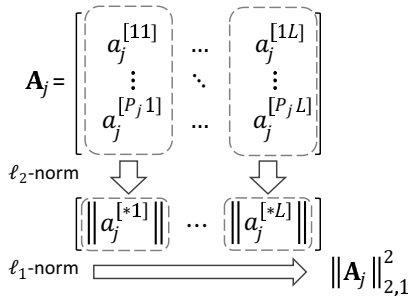


Figure 2: $\ell_{2,1}$-norm regularization for multi-task learning

A. Simulation Design

We considered simulation data of various complexities. The simulation data for training and testing were generated by extending the schemes of the previous related works [9], [14], [15]. For the simulation experimental setting, four-block and two-task input data were examined. Various dimensional random variables of an even number of samples for each block and task were generated. Four types of generation functions, which have different complexities and distributions, were designed as:

1) Type$_1(p, \mu^1, \mu^2)$: A p-dimensional random column vectors, which include data of two tasks, are generated from Gaussian distribution $\mathbf{x}^l = \mu^l + \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_p)$ and $l \in \{1, 2\}$.
2) Type$_2(p, \mu^1, \mu^2, \delta)$: A p-dimensional random column vectors $\mathbf{x}$ are generated by a random model with a threshold. The random model is $\mathbf{x}^l = \mu^l + \mathbf{1}_{u_{ij}} + \epsilon$, where u_{ij} is a uniform distributed random value and $\mathbf{1}_{u_{ij}} = 1$ if $u_{ij} \leq \delta$, otherwise 0 for $1 \leq i \leq N_l, 1 \leq j \leq p$.
3) Type$_3(p, \mu^1, \mu^2, \rho)$: This model generates the multi-

Table I: Simulation Data Generation

Simulation Data	Generation functions
$\mathbf{X}_1 = \{\mathbf{x}_{11}, \mathbf{x}_{12}, \mathbf{x}_{13}, \mathbf{x}_{14}\}$	$\mathbf{x}_{11} = \text{Type}_1(5, 2.4, -2.6)$
	$\mathbf{x}_{12} = \text{Type}_1(5, -2.6, 2.4)$
	$\mathbf{x}_{13} = \text{Type}_2(30, 1, -1, 0.7)$
	$\mathbf{x}_{14} = \text{Type}_3(60, 0, 0, 0.9)$
$\mathbf{X}_2 = \{\mathbf{x}_{21}, \mathbf{x}_{22}, \mathbf{x}_{23}, \mathbf{x}_{24}\}$	$\mathbf{x}_{21} = \text{Type}_1(5, -3, 4)$
	$\mathbf{x}_{22} = \text{Type}_1(5, 4, -3)$
	$\mathbf{x}_{23} = \text{Type}_3(50, 0, 0, 0.9)$
	$\mathbf{x}_{24} = \text{Type}_4(140, 2, 2, 2)$
$\mathbf{X}_3 = \{\mathbf{x}_{31}, \mathbf{x}_{32}, \mathbf{x}_{33}, \mathbf{x}_{34}\}$	$\mathbf{x}_{31} = \text{Type}_1(5, -5, 4)$
	$\mathbf{x}_{32} = \text{Type}_1(5, -3, 3)$
	$\mathbf{x}_{33} = \text{Type}_4(200, 0, 0, 1)$
	$\mathbf{x}_{34} = \text{Type}_3(90, 0, 0, 0.9)$

collinearity data which are highly correlated in more than two variables. The multicollinearity is often observed in biology data such as SNP. A $p \times N_l$ matrix samples are generated by multivariate normal distribution $\mathbf{x}^l \sim \mathcal{N}(\mu^l, \Sigma_{p\times p})$, where a covariance structure is derived by the first order of auto-regressive process, $\Sigma_{p\times p} = \{\sigma_{ij} | 1 \leq i, j \leq p\}$ and $\sigma_{ij} = \rho^{|i-j|}$.

4) $\text{Type}_4(p, \mu^1, \mu^2, \sigma)$: This model generates a normally distributed random variables, $\mathbf{x}^l \sim \mathcal{N}(\mu^l, \sigma\mathbf{I}_p)$.

The different numbers of variables of the multi-blocks (P_1 = 100, P_2 = 200, P_3 = 300, P_4 = 50), and of samples ($N_1 = N_2$ = 100, 200, or 500) were considered on the simulation study. For the first block data $\mathbf{X}_1$, the first five columns ($\mathbf{x}_{11}$) were generated by $\text{Type}_1(5, 2.4, -2.6)$, and the following five columns ($\mathbf{x}_{12}$) by $\text{Type}_1(5, -2.6, 2.4)$. The next 30 columns ($\mathbf{x}_{13}$) were produced by $\text{Type}_2(30, 1, -1, 0.7)$, and then the remaining columns ($\mathbf{x}_{14}$) were by $\text{Type}_3(60, 0, 0, 0.9)$. In a similar way to the first block, the other two blocks $\mathbf{X}_2$ and $\mathbf{X}_3$ were generated as described in Table. I.

The fourth block data $\mathbf{X}_4$ was generated by the linear model $\mathbf{X}_4^l = \sum_{j=1}^{3} b_j^l \mathbf{X}_j^l \mathbf{B}_j^l + \Xi$, where $\Xi \sim \mathcal{N}(0, \mathbf{I}_{P_j \times P_4})$. We assumed that $b_1^l = b_2^l = b_3^l = 1$ for simplicity. $\mathbf{B}_j^l$ ($1 \leq j \leq 3$) were designed so that the only first ten variables contribute to the first ten columns of $\mathbf{X}_4^l$. Let $\mathbf{0}_{\{k, P_i \times P_j\}}(\alpha)$ be a $P_i \times P_j$ dimensional matrix in which the first $k \times k$ block values are α and others are zeros. Thus, we defined that $\mathbf{B}_1^1 = \mathbf{0}_{\{10, P_1 \times P_4\}}(4.4), \mathbf{B}_1^2 = \mathbf{0}_{\{10, P_1\}}(-2.4), \mathbf{B}_2^1 = \mathbf{0}_{\{10, P_2\}}(3.4), \mathbf{B}_2^2 = \mathbf{0}_{\{10, P_2\}}(-2), \mathbf{B}_3^1 = \mathbf{0}_{\{10, P_3\}}(-3)$, and $\mathbf{B}_3^2 = \mathbf{0}_{\{10, P_3\}}(3)$.

We examined the performance of the proposed method **MBMTL** in terms of two views: (1) variable selection and (2) prediction performance. First, **MBMTL** is expected to identify significant factors of the integrative genomic model that represents complex disease mechanisms from the multi-block data. The variable selection performance is closely connected to a biomarker discovery problem, which is one of the most important research domains in the pharmaceutical industry. Secondly, the performance of a prognosis of gene expression level was tested in the proposed integrative genomic model when the input multi-block data such as **SNP**, **CNV**, and **DNA** methylation were given. The prediction of gene expression levels in the integrative genomic model enables computational biologists to conduct in *silico* experiments from the multi-block data.

The examination was compared with **SCCA** and **SGCCA** in 50 replications. **SCCA** [16] originally deals with two block data that consist of the multi-dimensional **X** and **Y** matrices by penalizing both ℓ_1 and ℓ_2 norm. The multi-block simulated data were combined into one $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, and considered $\mathbf{X}_4$ as a response matrix **Y** in order to directly compare **SCCA** with multi-block methods. For the multi-task setting, we performed **SCCA** separately for each task. **SGCCA** [10] is a sparse version of the generalized canonical correlation analysis that is a representative multi-block method. **SGCCA** also trained the model independently for each task as **SCCA**.

B. Variable Selection

The simulation data was designed so that only small subset variables of each block are significant in the multi-block data over the tasks. In this paper, only the first ten variables of each block were informative (non-zero coefficient) in the simulation setting. We built a confusion matrix in that setting:

1) True Positive (**TP**): correctly identified the informative features as significant
2) False Positive (**FP**): incorrectly identified uninformative features as significant
3) True Negative (**TN**): correctly identified uninformative features as insignificant
4) False Negative (**FN**): incorrectly identified informative features as insignificant.

Then, we measured True Positive Rate (**TPR**), Positive Predictive Value (**PPV**), True Negative Rate (**TNR**), and False Positive Rate (**FPR**) of the methods in over 50 replications. **TPR**, **PPV**, **TNR**, and **FPR** are defined as $\textbf{TPR} = \textbf{TP}/(\textbf{TP} + \textbf{FN})$, $\textbf{PPV} = \textbf{TP}/(\textbf{TP} + \textbf{FP})$, $\textbf{TNR} = \textbf{TN}/(\textbf{FP} + \textbf{TN})$, and $\textbf{FPR} = \textbf{FP}/(\textbf{FP} + \textbf{TN})$, respectively.

The experimental results for the variable selection are depicted in Fig. 3, and the detailed measurements of these experiments are listed in Table. II. **MBMTL** appeared as the best performance in the overall experiments. The simulation experiments assessed that **MBMTL** produces less **FP** and **FN** compared to **SCCA** and **SGCCA**.

C. Prediction Performance

For the assessment of the prediction performance, we measured Mean Squared Error (**MSE**) and absolute correlation (|**COR**|) between the latent variables of gene expression ($\mathbf{X}_4$) using 10-fold cross validation. The latent variable $\mathbf{v}_4$ can be derived from either the outer relation of Eq. (1) or the inner relation of Eq. (2). The latent variable from the outer relation is the linear combination of the other latent

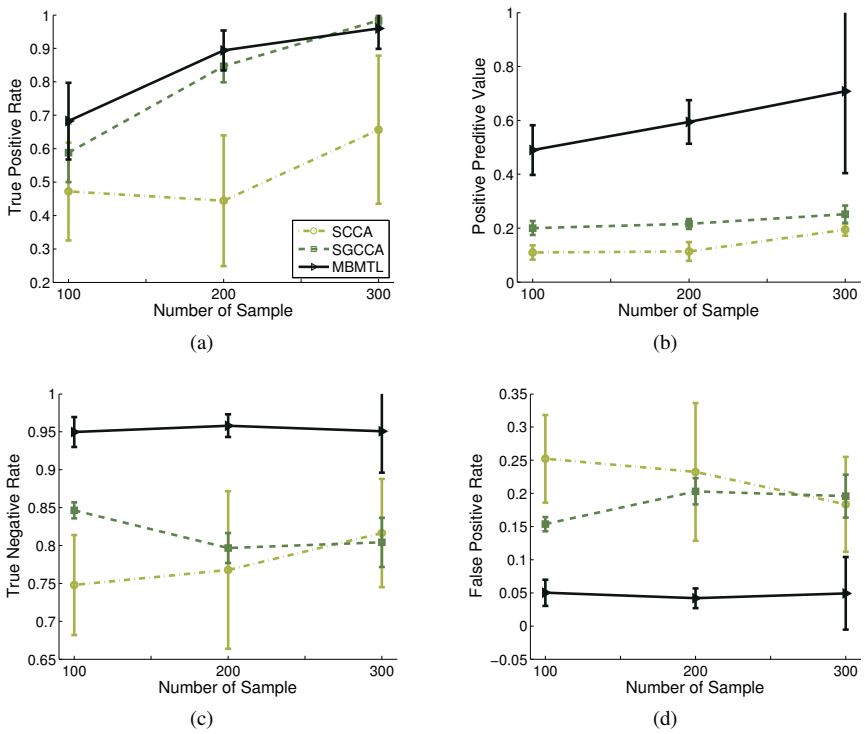


Figure 3: (a) True positive rate, (b) positive predictive value, (c) true negative rate, (d) false positive rate for variable selection performance assessment using simulated multi-block data.

Table II: Measurements for Variable Selection Results

	Samples	**SCCA**	**SGCCA**	**MBMTL**
TPR	100	0.47 ± 0.14	0.58 ± 0.08	0.68 ± 0.11
	200	0.44 ± 0.19	0.84 ± 0.04	0.89 ± 0.05
	500	0.65 ± 0.22	0.98 ± 0.01	0.96 ± 0.06
PPV	100	0.10 ± 0.02	0.20 ± 0.02	0.48 ± 0.09
	200	0.11 ± 0.02	0.21 ± 0.01	0.59 ± 0.08
	500	0.19 ± 0.09	0.25 ± 0.03	0.70 ± 0.30
TNR	100	0.74 ± 0.06	0.84 ± 0.01	0.94 ± 0.01
	200	0.76 ± 0.10	0.79 ± 0.01	0.95 ± 0.01
	500	0.81 ± 0.07	0.80 ± 0.03	0.95 ± 0.05
FPR	100	0.25 ± 0.06	0.15 ± 0.01	0.05 ± 0.01
	200	0.23 ± 0.10	0.20 ± 0.01	0.04 ± 0.01
	500	0.18 ± 0.07	0.19 ± 0.03	0.04 ± 0.05

variables of the multi-block data. For instance, when **SNP**, **CNV**, and **DNA** methylation data are given, we can predict gene expression levels in the integrative genomic model. For the experiment, **MBMTL** trained the model using the training multi-block data ($\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$). Then **MBMTL** produced the latent variable $\mathbf{v}_4$ using the test multi-block data for each task. **MSE** is computed by:

$$\textbf{MSE} = \sum_{l=1}^{L} (b_1^l \boldsymbol{\chi}_1^l \mathbf{a}_1^l + b_2^l \boldsymbol{\chi}_2^l \mathbf{a}_2^l + b_3^l \boldsymbol{\chi}_3^l \mathbf{a}_3^l - \boldsymbol{\chi}_4^l a_4^l)^2 / N_l, \quad (7)$$

where $\boldsymbol{\chi}_j$ is the j-th test block data. As a similar way, the absolute correlation $|\textbf{COR}|$ between the two latent variables

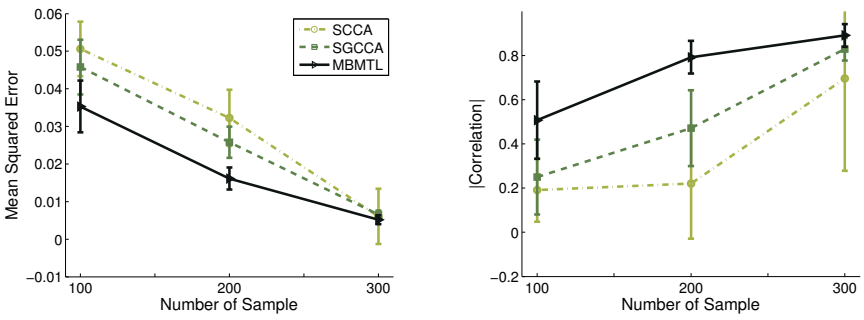
Table III: Measurements for Prediction Results

	Samples	**SCCA**	**SGCCA**	**MBMTL**
MSE	100	0.050 ± 0.007	0.045 ± 0.007	0.035 ± 0.006
	200	0.032 ± 0.007	0.025 ± 0.004	0.016 ± 0.002
	500	0.006 ± 0.007	0.006 ± 0.001	0.005 ± 0.001
|COR|	100	0.19 ± 0.14	0.24 ± 0.16	0.50 ± 0.17
	200	0.22 ± 0.25	0.47 ± 0.17	0.79 ± 0.07
	500	0.69 ± 0.41	0.82 ± 0.05	0.89 ± 0.05

were computed. The results of **MSE** and $|\textbf{COR}|$ are described in Fig. 4, and Table III. The outstanding performance of **MBMTL** was observed compared to **SCCA** and **SGCCA** in the simulation study.

IV. Human Brain Data of Psychiatric Disorders

We applied the proposed method **MBMTL** to human brain data of psychiatric disorders. Data from a total of 129 human brains were acquired from the human prefrontal cortex of three major psychiatric disorders such as schizophrenia (**SZ**), bipolar disorder (**BP**), major depression (**DP**) as well as from a control group. Specifically, the human brain data included 39 samples of **SZ**, 35 samples of **BP**, 12 samples of **DP**, and 43 samples of a control. From the sample, four genomic data of **SNP**, **CNV**, **DNA** methylation, and gene expression were considered for the



(a) (b)

Figure 4: (a) Mean squared error and (b) correlation for prediction performance assessment using simulated multi-block data

Table IV: Correlation of latent variables of the multi-blocks

	Correlation between	Gene expression
BP	**SNP**	0.912
	CNV	-0.800
	DNA methylation	-0.872
MP	**SNP**	-0.841
	CNV	-0.694
	DNA methylation	-0.835
SZ	**SNP**	0.822
	CNV	0.736
	DNA methylation	0.824
Control	**SNP**	0.893
	CNV	0.841
	DNA methylation	0.883

integrative genomic model. For a sample, 10,760 **SNP**s after removing highly correlated ones, 1,028 **CNV**s, 20,769 **DNA** methylation, and 19,767 gene expressions were examined. We considered four tasks using a disease label. We assumed that the psychiatric diseases share common features but have different functions.

As a result of the analysis using **MBMTL**, the integrative genomic association, which involves 38 **SNP**s, 19 **CNV**s, 196 **DNA** methylation, and 7 genes, was identified, where levels of high correlations between the subset of the multi-block data existed. The correlations between latent variables of the multi-block data in the psychiatric disorders and the control are described in Table. IV. The significantly high correlations between them ($0.694 \leq |\text{COR}| \leq 0.912$) supports the motivation and the justification of the integrative genomic model proposed. Then, we selected potential genetic markers of each multi-block data that significantly contributes to the integrative genomic model from the biological literature (Table. V). In the table, we list the markers derived from each block and their corresponding genes. The corresponding genes in **SNP**, **CNV**, and **DNA** methylation are derived by their locations. For instance, *rs7617733* is located within the *ALDH1L1* gene, and *rs17138705* is located in upstream of *ODZ4* (a.k.a. *TENM4*).

Among them, *ODZ4*, *PIP4K2A*, *NPAS3*, *NPY*, *CCKAR*, *BCL2*, *JAK2*, *RICS*, *ACADL*, *GRIN2A*, *LOX*, *PDGFA*, *CHRNA7*, *ASAHL*, *ADM*, *CCND2*, and *TPRKB* appeared in a large number of biological literature and *PubMed* database regarding cases on both **BP** disorder and **SZ**. For instance, **SNP**s within the gene *ODZ4* were reported as the genetic variants that are significantly associated to both **BP** and **SZ** in the previous genome-wide association studies [17]–[19]. The significant epigenetic changes of the **ACADL** gene in monozygotic twin of **BP** and **SZ** were also reported [20].

We also focused on the genetic factors which have different signs as well as comparable values of the coefficients between psychiatric disorders and a control. A **SNP**, *rs963300*, which is upstream of GNAO1, has a negative coefficient (-0.323) or almost zero (-4.7e-04) in the linear model of **BP** and **SZ** respectively; whereas the positive coefficient (1.265) in the control was shown. The data shows that the genetic variant of *rs963300* controls the changes of set of the gene expressions in **BP** and **SZ** less than in a control. Moreover, *CNP2061* made positive contributions (0.485 and 1.052) to the linear regression model of the **CNV** block in both **BP** and **SZ**, whereas it gave a negative (-0.523) contribution to the control. In this paper, we did not discuss major depression due to the small sample size. Nevertheless, the **MP** samples contributed to the common feature selection of psychiatric disorders as well as **BP** and **SZ**. The small size of **DP** is generally not considered by non-multi-task learning methods.

The gene regulatory network of the genes from the result were inferred by STRING database [21]. From the searched networks, the group where the most genes are connected to each other is depicted in Fig. 5. The inferred network shows that the discovered genes are highly co-related in similar gene regulatory networks or in related signaling pathways.

V. Conclusion

In this paper, we developed the novel multi-block and multi-task learning method in order to identify factors that play important roles in the integrative genomic model using the multiple genomic data of **SNP**, **CNV**, **DNA** methylation, gene expression. In addition to the capability of

Table V: The identified genetic markers of the multi-block data in psychiatric disorders

Source	ID	Coefficient				Chromosome	Corresponding Gene
		BP	DP	SZ	Control		
SNP	rs7617733	0.404	0.742	0.145	1.011	3	ALDH1L1
	rs17138705	0.338	-1.044	-0.211	0.338	11	Upstream of ODZ4
	rs963300	-0.323	0.162	-0.000	1.265	16	Upstream of GNAO1
	rs4953895	0.029	-0.212	0.075	0.128	2	Upstream of NAP5
	rs11208307	-0.074	-0.275	-0.017	0.142	1	ROR1
	rs2239265	0.355	0.018	-0.153	0.489	14	RGS6
	rs17668040	0.100	0.040	-0.098	0.212	22	Downstream of RP11-191L9.1
	rs7916957	0.249	-0.197	-0.141	-0.097	10	PIP4K2A
	rs17692624	0.365	-0.164	-0.115	0.215	8	MSRA
	rs10144775	0.841	0.668	-0.184	0.089	14	NPAS3
	rs17148810	0.032	-0.039	0.000	-0.009	7	Upstream of NPY
CNV	CNP975	-0.068	0.076	-1.139	0.353	10	PARG
	CNP2061	0.485	-0.056	1.052	-0.523	6	HLA-DRB5
	CNP10045	0.294	0.000	0.000	0.598	1	CROCC
DNA methylation	cg15928132	-0.022	-0.216	0.030	-0.178	4	CCKAR
	cg01803238	-0.057	0.085	0.083	-0.199	18	BCL2
	cg09494188	-0.019	-0.003	0.054	-0.156	11	SCGB1A1
	cg20394284	-0.087	0.110	0.094	0.074	9	JAK2
	cg15583072	-0.061	0.204	-0.005	0.041	12	IAPP
	cg25978208	0.041	-0.113	0.036	-0.019	15	SNRPN
	cg23268677	-0.040	0.033	0.072	0.032	22	TXNRD2
	cg25017304	-0.035	-0.038	0.015	0.046	1	ARHGEF11
	cg20892287	-0.014	0.060	-0.004	0.009	11	RICS
	cg18515587	0.025	0.051	0.009	0.017	1	SELENBP1
	cg14795968	-0.076	0.045	-0.040	-0.037	2	ACADL
	cg04041960	-0.137	0.026	0.033	0.032	10	RGS10
	cg17428423	0.001	0.023	-0.008	-0.014	16	DOC2A
	cg13948987	-0.003	-0.036	0.000	0.007	6	HCRTR2
	cg15604467	-0.039	0.061	0.001	-0.009	13	POU4F1
	cg25047001	-0.153	-0.076	-0.055	-0.052	16	GRIN2A
	cg01824804	-0.076	-0.009	-0.040	-0.056	5	LOX
	cg22476295	-0.009	0.002	-0.004	0.017	7	PDGFA
	cg20861607	0.000	-0.008	-0.002	0.004	15	CHRNA7
	cg03201604	-0.001	-0.003	-0.002	0.002	4	ASAHL
	cg12228707	0.017	-0.001	-0.005	-0.018	11	ADM
	cg16719404	-0.003	0.003	-0.003	0.002	1	CD2
	cg22511947	0.008	-0.003	0.006	0.002	2	FN1
	cg25268451	-0.005	0.001	0.003	-0.003	20	GNAS
	cg16310717	-0.005	-0.003	0.001	-0.003	12	CCND2
	cg22338307	-0.059	0.027	0.024	0.007	2	TPRKB
Gene	7963664	2.948	3.462	-2.293	3.227	12	SP7
	8037513	1.137	0.764	1.412	1.420	19	EXOC3L2
	8009705	1.476	1.938	0.156	1.793	17	OTOP2
	8144577	0.748	0.562	-0.541	0.512	8	C8orf74
	8003804	0.614	0.451	0.124	0.530	17	OR1A1
	7906622	1.199	0.946	0.193	0.503	1	LY9
	8111524	0.173	0.107	0.024	0.115	5	UGT3A2

MBMTL's multi-block analysis, **MBMTL** solves a multi-task problem where the multi-tasks share common features but have different functions. The settings of both multi-block and multi-task is essential in the integrative genomic study that examines integrated mechanism of human diseases and identifies biomarkers that cause the diseases. We first built an integrative genomic model using **SNP**, **CNV**, **DNA** methylation, and gene expression data. Then, we performed the proposed **MBMTL** for the integrative genomic model of the psychiatric disorders and a control. **MBMTL** identified subsets of genetic markers that are associated to the variants of the gene expression from the multi-block data. We discussed the findings from the multi-block and multi-task analysis on the psychiatric disorders by biological literature and the interaction network studies. The proposed method **MBMTL** can be easily extended by introducing new types of genomic data in multiple diseases.

REFERENCES

[1] J. Hirschhorn and M. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, pp. 95–108, 2005.

[2] Y. Gilad, S. Rifkin, and J. Pritchard, "Revealing the architecture of gene regulation: the promise of eQTL studies," *Trends in Genetics*, vol. 24, pp. 408–415, 2008.

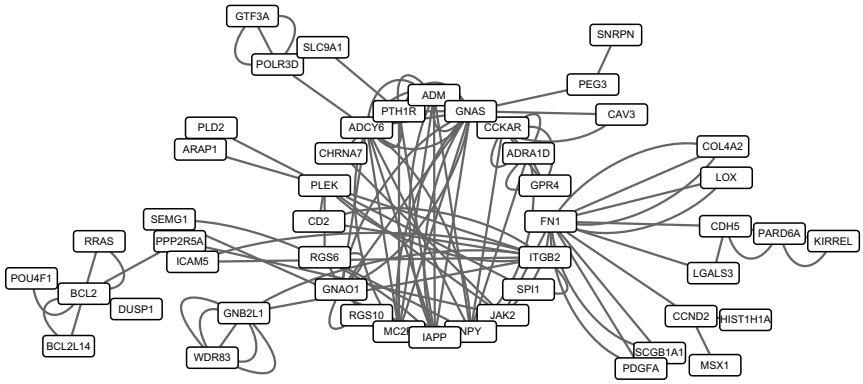


Figure 5: The gene regulatory network searched with the gene results by STRING database.

[3] J. L. Freeman *et al.*, "Copy number variation: New insights in genome diversity," *Genome Research*, vol. 16, no. 8, pp. 949–961, 2006.

[4] S. Girirajan, C. D. Campbell, and E. E. Eichler, "Human Copy Number Variation and Complex Genetic Disease," *Annual Review of Genetics*, vol. 45, no. 1, pp. 203–226, 2011.

[5] E. N. Gal-Yam, Y. Saito, G. Egger, and P. A. Jones, "Cancer Epigenetics: Modifications, Screening, and Therapy," *Annual Review of Medicine*, vol. 59, no. 1, pp. 267–280, 2008.

[6] L. D. Moore, T. Le, and G. Fan, "DNA Methylation and Its Basic Function," *Neuropsychopharmacology*, vol. 38, pp. 23–38, 2013.

[7] T. C. G. A. R. Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[8] J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen1, and M. Blanchette, "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome Biology*, vol. 15, p. R37, 2014.

[9] W. Li, S. Zhang, C. Liu, and X. Zhou, "Identifying multi-layer gene regulatory modules from multi-dimensional genomic data," *Bioinformatics*, vol. 28, no. 19, pp. 2458–2466, 2012.

[10] M. Kang, B. Zhang, X. Wu, C. Liu, and J. Gao, "Sparse generalized canonical correlation analysis for biological model integration: a genetic study of psychiatric disorders," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, July 2013, pp. 1490–1493.

[11] Q. Zhao, X. Shi, Y. Xie, J. Huang, B. Shia, and S. Ma, "Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA," *Briefings in Bioinformatics*, 2014.

[12] A. Lykou and J. Whittaker, "Sparse CCA using a Lasso with positivity constraints," *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3144–3157, dec 2010.

[13] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 41–48.

[14] K. L Cao, D. Rossouw, C. Robert-Grani, and P. Besse, "A Sparse PLS for Variable Selection when Integrating Omics Data," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, 2008.

[15] H. Chun and S. Keles, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society*, vol. 72, pp. 3–25, 2010.

[16] S. Waaijenborg and A. H. Zwinderman, "Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers," in *Proceedings 1(Suppl 1):S122*, December 2007.

[17] N. Craddock and P. Sklar, "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4," *Nature Genetics*, vol. 43, pp. 977–983, 2011.

[18] N. Craddock and P. Sklar, "Genetics of bipolar disorder," *The Lancet*, vol. 381, pp. 1654–1662, 2013.

[19] T. Muhleisen *et al.*, "Genome-wide association study reveals two new risk loci for bipolar disorder," *Nature Communications*, vol. 5, 2014.

[20] D. EL *et al.*, " Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder," *Human Molecular Genetics*, vol. 20, pp. 4786–4796, 2011.

[21] A. Franceschini *et al.*, "String v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. D1, pp. D808–D815, 2013.