Published in final edited form as:

Int J Data Min Bioinform. 2017 ; 18(3): 223–239. doi:10.1504/IJDMB.2017.10008266.

Integration of multi-omics data for integrative gene regulatory network inference

Neda Zarayeneh,

Department of Computer Science, Texas A&M University Commerce, Commerce, TX, USA

Euiseong Ko,

Department of Computer Science, Kennesaw State University, Marietta, GA, USA

Jung Hun Oh,

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Sang Suh,

Department of Computer Science, Texas A&M University Commerce, Commerce, TX, USA

Chunyu Liu,

Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA

Jean Gao,

Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA

Donghyun Kim, and Mingon Kang

Department of Computer Science, Kennesaw State University, Marietta, GA, USA

Abstract

Gene regulatory networks provide comprehensive insights and indepth understanding of complex biological processes. The molecular interactions of gene regulatory networks are inferred from a single type of genomic data, e.g., gene expression data in most research. However, gene expression is a product of sequential interactions of multiple biological processes, such as DNA sequence variations, copy number variations, histone modifications, transcription factors, and DNA methylations. The recent rapid advances of high-throughput omics technologies enable one to measure multiple types of omics data, called ‘multi-omics data’, that represent the various biological processes. In this paper, we propose an Integrative Gene Regulatory Network inference method (iGRN) that incorporates multi-omics data and their interactions in gene regulatory networks. In addition to gene expressions, copy number variations and DNA methylations were considered for multi-omics data in this paper. The intensive experiments were carried out with simulation data, where iGRN’s capability that infers the integrative gene regulatory network is assessed. Through the experiments, iGRN shows its better performance on model representation and interpretation than other integrative methods in gene regulatory network inference. iGRN was

Correspondence to: Mingon Kang.

This paper is a revised and expanded version of a paper entitled ‘Integrative gene regulatory network inference using multi-omics data’ presented at the ‘IEEE International Conference on Bioinformatics & Biomedicine (IEEE BIBM)’, Shenzhen, China, 15–18 December 2016.

also applied to a human brain dataset of psychiatric disorders, and the biological network of psychiatric disorders was analysed.

Keywords

gene regulatory network inference; multi-omics data; data integration

1 Introduction

Gene regulatory networks (GRNs) provide comprehensive insights and in-depth understanding of complex biological processes (Chai et al., 2014; Hill et al., 2016). GRNs describe molecular interactions of complex biological processes by typically using a graph model, where nodes and edges represent genes and their regulations respectively. However, gene regulatory network inference, reconstructing transcriptional interactions between gene from microarray expression, is challenging due to the extremely large scale of microarray gene expression data. Moreover, the microarray gene expression data are typically noisy, and the number of available samples is much less than the potential interactions in the gene regulatory network.

The importance of an integrative genomic study is being increasingly emphasised along with the rapid development of various types of high-throughput genomic data. Biological systems involve a sequence of complex interactions in multiple biological processes such as genetic, epigenetic, and transcriptional regulation (Kristensen, 2014; Higdon, 2015; Rebollar, 2016; Cisek et al., 2015). Specifically, gene expression is a product of sequential interactions of single nucleotide polymorphism (SNP), copy number variation (CNV), histone modifications, transcription factor, DNA methylation, other genes in relevant pathways, and many other factors (Aure et al., 2013; Wagner et al., 2014; Kang et al., 2015; Kim et al., 2015; Kang et al., 2016).

Therefore, the multiple types of biological data, called *multi-omics data*, should be considered to explain such complex biological processes. In this paper, we consider CNVs and DNA methylations as well as gene expression data for gene regulatory network inference. CNV, which is a modified gene structure, often alters downstream pathways or regulatory networks. DNA methylation also often reduces gene expression in a nearby gene by the methyl groups added to the DNA. For instance, monozygotic twins have an identical DNA sequence. However, they often show discordance in congenital heart defects due to the discrepancy in CNVs (Breckpot et al., 2012; Henrichsen et al., 2009).

The approaches of gene regulatory network inference are mainly four-fold: (1) correlation-based, (2) Bayesian-based, (3) regression-based, and (4) integrative approaches. Correlation-based approaches identify the interactions between genes by using linear correlation (e.g., Pearson's correlation coefficient) on Boolean networks, which is the simplest model of gene regulatory networks. Boolean networks represent an active interaction as one and inactive as zero, and then active interactions of genes are determined if the correlation coefficient is higher than a certain threshold. WGCNA (Weighted Gene Coexpression Network Analysis) is one of the most widely used correlation-based methods (Langfelder and Hovath, 2008).

Along with the correlation coefficients, mutual information (MI) (Atlay and Emmert-Streib, 2010), maximum information correlation (MIC) (Reshef et al., 2011), and conditional mutual information (CMI) (Liang and Wang, 2008; Zhang et al., 2012) have been also employed on Boolean networks for gene regulatory network inference. However, the correlation-based approaches lack the interpretation of causal effect between genes since they use an undirected graph.

Bayesian-based approaches for gene regulatory networks allow one to infer casual relationships between genes based on a directed acyclic graph (DAG) in a probabilistic manner (Njah and Jamoussi, 2015; Young et al., 2014; Werhli and Husmeier, 2007; Husmeier and Werhli, 2007). Moreover, Bayesian network inference approaches can deal with the missing value problem and can efficiently incorporate prior knowledge such as pathway databases in the Bayesian model (Werhli and Husmeier, 2007; Husmeier and Werhli, 2007). However, it limits the number of genes to infer the gene regulatory network, since the learning cost grows exponentially.

Regression-based approaches infer the relationships of genes by decomposing the gene regulatory networks into P (number of genes) regression problems (Omranian et al., 2016; Oh and Deasy, 2014; Gustafsson et al., 2005). The regression-based approaches have often adapted the Least Absolute Shrinkage and Selection Operator method (LASSO) for providing biologically interpretable solutions on the decomposed regression models (Tibshirani, 1996). Such approaches showed robust performance even with large-scale gene expression data. Regression-based approaches were highly ranked as top-performing methods in the HPN-DREAM network inference challenge (Hill et al., 2016).

A number of integrative research has been performed for gene regulatory network inference. SNP-gene regulatory network (SGRN) was proposed to incorporate SNPs as regulators of genes in the gene regulatory networks (Cai et al., 2013). The sparse Structural Equation Models (SEM) for SGRN enables one to identify cis-expression quantitative trait loci (cis-eQTL) in a SGRN model. DCGRN was developed as an integrative gene regulatory network inference method with multi-omics data, such as gene expression, CNV, and DNA methylation (Kim et al., 2014). In SGRN and DCGRN, the multiple types of omics data of SNPs, DNA methylations, CNVs and gene expressions are considered as nodes in the directed network models, and their significant interactions are identified by LASSO. A number of biological databases, such as protein-protein interactions, transcription factor DNA-binding, and gene knock-down, are integrated to produce robust gene regulatory networks by using random forest (Petràlia et al., 2015).

In this paper, we propose anovel Integrative Gene Regulatory Network inference (iGRN) method. The contributions of this study are:

- to infer a gene regulatory network by incorporating multi-omics data such as CNV, DNA methylation, and gene expression, while most gene regulatory network inference methods used a single platform of gene expression, and
- to biologically interpret the transcriptional regulations between genes by taking into account the interaction effects between the multi-omics data.

We first define notations in Section 2.1, and then describe our proposed method iGRN in Section 2.2. In Section 3, we show the intensive experimental results with simulation data and compare the performance with current state-of-the-art methods. Finally, we apply iGRN to human brain data of psychiatric disorders and analyse the gene regulatory network inferred by the proposed method in Section 4.

2 Methods

2.1 Notations

We describe the notations used throughout this paper. Let $\mathbf{G}$ denote a matrix of gene expression, $\mathbf{G} \in \mathbb{R}^{N \times P}$, where N is the number of the samples and P is the number of genes in microarray data. The vector of gene expression at the i^{th} gene is denoted as $\mathbf{g}_i \in \mathbb{R}^N$, and $\mathbf{G}_{-i}$ represents the matrix that contains gene expressions other than gene i , $\mathbf{G}_{-i} \in \mathbb{R}^{N \times (P-1)}$. The matrices of CNV and DNA methylation data are denoted as $\mathbf{C} \in \mathbb{R}^{N \times V}$ and $\mathbf{D} \in \mathbb{R}^{N \times M}$, respectively. We suppose that CNVs or DNA methylations can be annotated to nearby genes (upstream or downstream of the gene) in the same chromosome. The gene annotations of the CNV and DNA methylation to gene i are represented as the matrices of $\mathbf{C}_{\{i\}} \in \mathbb{R}^{N \times V_{\{i\}}}$ and $\mathbf{D}_{\{i\}} \in \mathbb{R}^{N \times M_{\{i\}}}$ respectively, where $V_{\{i\}}$ and $M_{\{i\}}$ are the numbers of CNVs and DNA methylations which are annotated to gene i .

The regulatory relationships between genes are represented by an adjacency matrix of gene expression $\mathbf{B} \in \mathbb{R}^{P \times P}$, and integrative interactions of multi-omics data other than gene expression are expressed by their own biadjacency matrices. In iGRN, the interactions of CNVs and DNA methylations to genes are described as $\mathbf{B}_C \in \mathbb{R}^{V \times P}$ and $\mathbf{B}_D \in \mathbb{R}^{M \times P}$ respectively. We assume that there is no self-regulation in the gene regulatory network, i.e., $\mathbf{B}_{ii}=0$ where $i = \{1, \dots, P\}$.

2.2 Integrative gene regulatory network inference

We propose an Integrative Gene Regulatory Network inference (iGRN) method that infers a gene regulatory network from multi-omics data. The current state-of-the-art methods for integrative gene regulatory network inference using multi-omics data, such as SGRN (Cai et al., 2013) and DCGRN (Kim et al., 2014), consider all the types of data as nodes in networks. In other words, nodes can indicate genes, CNVs, or DNA methylations. In contrast, iGRN constructs homogeneous gene regulatory networks where nodes represent only genes, which consequently makes it possible to apply most graph algorithms for further analysis. Figure 1 shows a simple integrative gene regulatory network, where a gene (G1) regulates another gene (G4) with biological processes of a CNV (CNV2) and a DNA methylation (DM2).

The proposed method, iGRN, represents integrative gene regulatory networks with multi-layered adjacency matrices of the multi-omics data. It constructs the adjacency matrix of gene expression and the biadjacency matrices of CNV and DNA methylation. The adjacency matrix of gene expression defines the basic structure of the transcriptional biological networks, and the biadjacency matrices of CNV and DNA methylation describe their integrative interactions on the gene regulations.

For formulating the integration of the heterogeneous data into a standardised format, iGRN takes into account the interaction effects of CNVs and DNA methylations with genes. The integrative interactions between a gene i and its nearby CNVs and DNA methylations can be described by Fisher's interaction model:

$$\mathbf{g}_i \otimes \mathbf{C}_{\{i\}}, \quad \mathbf{g}_i \otimes \mathbf{D}_{\{i\}},$$

where $\otimes$ is an element-by-element multiplication. It explains different gene expression levels on the variations of CNVs or DNA methylations.

Thus, the expression of gene i can be represented by a sparse linear model by incorporating not only other genes but also interaction effects of its nearby CNVs and DNA methylations. The gene expression ($\mathbf{g}_i$) for gene i is formulated by:

$$\mathbf{g}_i = \mathbf{G}_{-i} \mathbf{b}_i^g + (\mathbf{C}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^c + (\mathbf{D}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^d + \varepsilon_i, \text{ subject to } |\mathbf{b}_i^g| \leq C_g, |\mathbf{b}_i^c| \leq C_c, |\mathbf{b}_i^d| \leq C_d, \quad (2)$$

where $\mathbf{b}_i^g$, $\mathbf{b}_i^c$ and $\mathbf{b}_i^d$ are the coefficients of gene expressions other than gene i , CNVs and DNA methylations of gene i , respectively. $|\cdot|$ is the L-1 norm, and the residual is denoted as ε_i . The adjacency matrix $\mathbf{B}$ of the gene regulatory network is comprised of $\mathbf{b}_i^g$ ($1 \leq i \leq P$) in (2), i.e., $\mathbf{B} = \{\mathbf{b}_1^g, \dots, \mathbf{b}_P^g\}^\perp$.

The biadjacency matrices of CNVs $\mathbf{B}_C$ and DNA methylations $\mathbf{B}_D$ are also constructed by $\mathbf{b}_i^c$ and $\mathbf{b}_i^d$.

The integrative gene regulatory network can be inferred by optimising the parameters of (2).

The learning function $F(\mathbf{b}_i^g, \mathbf{b}_i^c, \mathbf{b}_i^d)$ for the optimal parameters is obtained by using least squares with the sparse setting:

$$\operatorname{argmin} F(\mathbf{b}_i^g, \mathbf{b}_i^c, \mathbf{b}_i^d) = \|\mathbf{g}_i - (\mathbf{G}_{-i} \mathbf{b}_i^g + (\mathbf{C}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^c + (\mathbf{D}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^d)\|^2 + \lambda_g |\mathbf{b}_i^g| + \lambda_c |\mathbf{b}_i^c| + \lambda_d |\mathbf{b}_i^d|, \quad (3)$$

where λ_g , λ_c and λ_d are hyper-parameters for sparsity regularisation, and $|\cdot|^2$ is the L-2 norm. The optimisation function can be considered as the following LASSO problem:

$$\operatorname{argmin} \|\mathbf{g}_i - \mathbf{X} \mathbf{b}_i\|^2 + \lambda |\mathbf{b}_i|, \quad (4)$$

where $\mathbf{X}$ is the augmented matrix, $\mathbf{X} = \{\mathbf{G}_{-i}, \mathbf{C}_{\{i\}} \otimes \mathbf{g}_i, \mathbf{D}_{\{i\}} \otimes \mathbf{g}_i\}$. However, the number of genes in $(P-1)$ is much larger than of CNVs ($V_{\{i\}}$) and DNA methylations ($M_{\{i\}}$) associated to gene i . For instance, there are only a couple of CNVs ($\mathbf{C}_{\{i\}}$) or DNA methylations

$(\mathbf{D}_{\{i\}})$ for a gene in the psychiatric disorder data that we used for the experiment in the paper, whereas the number of genes in $(\mathbf{G}_{-i})$ is hundreds even after pre-processing. Thus, the solution of LASSO may tend to ignore most CNVs and DNA methylations despite their importance.

Therefore, we solve the optimisation problem in a stepwise manner. First, we identify significant genes that interact with gene i from $\mathbf{G}_{-i}$ by LASSO:

$$\operatorname{argmin} \|\mathbf{g}_i - \mathbf{G}_{-i} \mathbf{b}_i^g\|^2 + \lambda \|\mathbf{b}_i^g\|. \quad (5)$$

The matrix of $\mathbf{G}_{-i}'$ is constructed with the genes with non-zero coefficients. Secondly, we compute p-values of the variables in the following linear regression:

$$\mathbf{g}_i = \mathbf{G}_{-i}' \mathbf{b}_i^{g'} + (\mathbf{C}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^c + (\mathbf{D}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^d + \varepsilon_i. \quad (6)$$

The coefficients of the genes, CNVs, and DNA methylations with p-values > 0.05 are set to zeros. Then, the coefficients for genes are assigned to the adjacency matrix, and $\mathbf{b}_i^c$ and $\mathbf{b}_i^d$ are assigned to the biadjacency matrices of $\mathbf{B}_C$ and $\mathbf{B}_D$ respectively. The procedure is described in Algorithm 1.

Algorithm 1

```

1: for  $i \in \{1, \dots, P\}$  do
2:    $\mathbf{b}_i^g = \text{LASSO}(\mathbf{G}_{-i}, \mathbf{g}_i)$ 
3:   Compute the linear regression of (6)
4:    $\mathbf{b}_{ij}^{g'} = \begin{cases} \mathbf{b}_{ij}^{g'} & \text{if } \mathbf{b}_{ij}^{g'} \text{ is non-zero and } p\text{-value}(\mathbf{b}_{ij}^{g'}) < 0.05 \\ 0 & \text{otherwise} \end{cases}$ 
5:    $\mathbf{b}_{ij}^c = \begin{cases} \mathbf{b}_{ij}^c & \text{if } \mathbf{b}_{ij}^c \text{ is non-zero and } p\text{-value}(\mathbf{b}_{ij}^c) < 0.05 \\ 0 & \text{otherwise} \end{cases}$ 
6:    $\mathbf{b}_{ij}^d = \begin{cases} \mathbf{b}_{ij}^d & \text{if } \mathbf{b}_{ij}^d \text{ is non-zero and } p\text{-value}(\mathbf{b}_{ij}^d) < 0.05 \\ 0 & \text{otherwise} \end{cases}$ 
7:   Construct  $\mathbf{B}, \mathbf{B}_C$ , and  $\mathbf{B}_D$ 
8: end for

```

3 Simulation studies

We conducted intensive simulation experiments to evaluate our proposed method and compare the performance with existing methods. Due to only few available well-known true models of biological networks, the assessment of gene regulatory network inference in

complex organisms such as human is challenging. Thus, the performance was indirectly evaluated with simulation data that implements integrative biological networks where the true model is given.

We generated the simulation data under the assumption that we hypothesised for the integrative gene regulatory networks. In the simulation studies, we aim to (1) verify that our proposed method produces robust performance to identify the true models of gene regulatory networks from multi-omics data, and (2) to compare the performance with current state-of-the-art methods on the given hypothesis. We carried out the following three experiments with the simulation data: (1) Receiver Operating Characteristic (ROC) curve, (2) sensitivity, and (3) false discovery rate.

3.1 Simulation settings

In the integrative gene regulatory network model, gene expression can be represented by two components: (1) gene expression regulated by other genes ($\mathbf{G}^g$) and (2) interactions of CNVs and DNA methylations ($\mathbf{G}^i$), as shown in (2):

$$\mathbf{G} = \mathbf{G}^g + \mathbf{G}^i, \quad (7)$$

where

$$\begin{aligned} \mathbf{G}^g &= \mathbf{G}_{-i} \mathbf{b}_i^g, \\ \mathbf{G}^i &= (\mathbf{C}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^c + (\mathbf{D}_{\{i\}} \otimes \mathbf{g}_i) \mathbf{b}_i^d. \end{aligned}$$

First, $\mathbf{G}^g$ was generated by the given adjacency matrix $\mathbf{Z}$:

$$\mathbf{G}^g = \mathbf{E}(\mathbf{I} - \mathbf{Z})^{-1}, \quad (8)$$

where $\mathbf{I} \in \mathbb{R}^{N \times P}$ is an identity matrix, and $\mathbf{E} \in \mathbb{R}^{N \times P}$ is a matrix with normally distributed random values for noise, $\mathbf{E} \sim N(0, 0.01)$. The adjacency matrix $\mathbf{Z}$ is a sparse acyclic graph without self-loop.

The CNV data ($\mathbf{C} \in \mathbb{R}^{N \times P}$) was implemented by taking the values {0,1, 2, 3,4} with the corresponding probabilities {0.01, 0.02, 0.4, 0.2, 0.1}. The given probabilities were directly acquired from CNV of human brain data that we have used in Section 4. The DNA methylation ($\mathbf{D} \in \mathbb{R}^{N \times M}$) was randomly obtained by the uniform distribution on the interval [0,1]. In practice, CNVs and DNA methylations were annotated to nearby genes by using their loci and gene regions. We designated the associations by sparse Boolean mapping matrices $\mathbf{W} \in \mathbb{R}^{V \times P}$ and $\mathbf{F} \in \mathbb{R}^{M \times P}$ for CNVs and DNA methylations, where only a couple of CNVs and DNA methylations can be annotated to a gene. In this simulation data, we assume that all of the CNVs and DNA methylations nearby a gene significantly regulate the gene expression.

The gene expression regulated by the interactions of CNVs and DNA methylations was generated by:

$$\mathbf{G}^i = \mathbf{C}\mathbf{W} \otimes \mathbf{G} + \mathbf{D}\mathbf{F} \otimes \mathbf{G}. \quad (9)$$

The gene expression controls the gene expression levels of other genes with the interaction effects of multi-omics data in gene regulatory networks. Therefore, we repeated (8) and (9) until $\mathbf{G}$ converges. Note that $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{F}$ are the (bi)adjacency matrices of ground truth in the simulation studies. The algorithm is described in Algorithm 2.

Algorithm 2

```

1:  $\mathbf{G} = \mathbf{E}(\mathbf{I} - \mathbf{Z})^{-1}$ 
2: do
3:    $\mathbf{G} = (\mathbf{E} + \mathbf{C}\mathbf{W} \otimes \mathbf{G} + \mathbf{D}\mathbf{F} \otimes \mathbf{G})(\mathbf{I} - \mathbf{Z})^{-1}$ 
4: while  $\mathbf{G}$  converges

```

We considered a LASSO-based GRN method (GRN) as baseline and DCGRN (Kim et al., 2014) which is an integrative gene regulatory network inference method that uses multi-omics data. GRN infers the gene regulatory relationship on gene i with LASSO regularisation:

$$\mathbf{g}_i = \mathbf{G}_{-i} \mathbf{b}_i^g + \varepsilon_i, \text{ subject to } |\mathbf{b}_i^g| < C_g. \quad (10)$$

GRN identifies significant gene regulations by LASSO solution, but it considers only gene expression data for the network inference. In contrast, DCGRN incorporates multi-omics data of CNVs and DNA methylations in the model:

$$\mathbf{g}_i = \mathbf{G}_{-i} \mathbf{b}_i^g + C_{(i)} \mathbf{b}_i^c + \mathbf{D}_{(i)} \mathbf{b}_i^d + \varepsilon_i, \text{ subject to } |\mathbf{b}_i^g| \leq C_g, |\mathbf{b}_i^c| \leq C_c, \text{ and } |\mathbf{b}_i^d| \leq C_d. \quad (11)$$

3.2 Experimental results with simulation data

First, we evaluated the performance by computing the area under the receiver operating characteristic curve (AUROC). The confusion matrix of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) is defined as:

- TP: correctly identified the positive gene regulations as non-zero coefficients,
- FP: incorrectly identified the positive gene regulations as zero coefficients,
- TN: correctly identified the negative gene regulations as zero coefficients,
- FN: incorrectly identified the negative gene regulations as non-zero coefficients.

The non-zero coefficients of b_i^g , b_i^d , and b_i^c were considered as positives, while the coefficients of zero were negatives. The confusion matrices for gene regulations and integrative interactions of CNVs and DNA methylations were separately computed.

The ROC curves were traced over different thresholds to examine the trade-off between True Positive Rate ($TPR = TP / (TP + FN)$) and False Positive Rate ($FPR = FP / (FP + TN)$). The hyper-parameters (λ_g , λ_c , and λ_d) in (3) determine the sparsity of significant components with non-zero coefficients in the multi-omics data. Note that all of the coefficients are non-zero when the parameter is zero, while all coefficient values become zero when infinite value is given for the parameter. We considered the sparsity step ($1 - \theta = P / (P + V + M)$) that determines the hyper-parameters in LASSO solution. In this simulation study for the ROC curves, only the coefficient values were considered to determine the positive interactions, where p-values were not computed.

GRN computes only a confusion matrix for gene regulations, while DCGRN and iGRN have confusion matrices for CNVs and DNA methylations as well as gene expression. Therefore, overall ROC curves were considered, where only the confusion matrix of gene regulation was reflected on GRN, while the three confusion matrices were combined to compute ROC curves in DCGRN and iGRN. The overall ROC curves are illustrated in Figure 2, and AUROC is shown in Table 1. The experimental result of the overall AUROC supports that iGRN (0.938) provides better performance than GRN (0.895) and DCGRN (0.843).

TPRs on interactions of the CNVs and DNA methylations were measured for DCGRN and iGRN. Since the simulation data does not include negatives on CNVs and DNA methylations, we examined how well the methods identify the true positives. The TPRs are shown in Figure 3, where iGRN outperforms DCGRN in identifying true integrative interactions of CNVs and DNA methylations.

Secondly, we measured the overall sensitivity which is the probability of identifying the true positives. In this simulation study, the hyper-parameters were optimised by 10-fold cross-validation. The multi-omics elements with non-zero coefficients and whose p-values are less than 0.05 are considered as positives. The overall sensitivity is depicted in Figure 4. iGRN produced the best sensitivity (0.300 ± 0.034), while GRN and DCGRN showed 0.199 ± 0.030 and 0.269 ± 0.035 respectively. The sensitivities of CNVs and DNA methylations on iGRN and DCGRN are shown in Figure 5. The sensitivities for iGRN and DCGRN were 0.102 ± 0.035 and 0.054 ± 0.030 , respectively.

Lastly, we conducted the simulation study for False Discovery Rate (FDR). In this study, we generated the simulation data that have no gene-gene regulation in the biological network. All positive predictions that the methods infer are false positives, since the true adjacency matrix has all zero. FDR is computed as $FP / (TP + FP)$. The FDRs of GRN, DCGRN, and iGRN less than 0.02 are observed in Figure 6, where the FDRs were 0.019 ± 0.003 , 0.019 ± 0.003 , 0.019 ± 0.003 respectively. It shows that iGRN has a chance less than 2% that misidentifies interactions.

4 Human brain data for psychiatric diseases

We performed iGRN on a human brain dataset that consists of psychiatric diseases and control with 131 samples. The multi-omics data of gene expression, CNV, and DNA methylation used in the preparation of this paper were obtained from the human prefrontal cortex (Liu et al., 2010). The human brain data contains 39 samples of schizophrenia, 35 of bipolar disorder, 12 of major depression patients, and 44 of healthy control samples, where each sample has 25,833 of gene expression measurement, 1,028 of CNV, and 24,399 of DNA methylation. We considered the psychiatric disorder data as a group combining the three psychiatric disorder samples, since the psychiatric disorders share many common biological features.

We examined only 495 genes among the 25,833 gene expression data. Fold changes (FC) and p-values of the two groups (psychiatric diseases and healthy control) were computed, and the 495 genes, where their p-values < 0.05 and ($FC < 0.09$ or $FC > 1.1$), were selected for further analysis (see Figure 7). Then, we annotated CNVs and DNA methylations based on their locations in the chromosome. A CNV is annotated to gene i if its region is overlapped with gene i in the same chromosome. DNA methylation is annotated to gene i if its locus is located within 2Kbp around the transcription start site (TSS) of gene i .

The integrative gene regulatory network inferred by iGRN is depicted in Figure 8. The 495 genes were introduced to the String database (<http://string-db.org>) (Szklarczyk et al., 2015), and the gene interactions were compared. Among the 495 genes, 190 genes (coloured in red in Figure 8) were reported in the String database, and 41 interactions were matched. The simplified sub-components that maximise interaction coefficients of the connected-components were analysed (Ko et al., 2017) for further analysis, as shown in Figure 9. Two gene regulations, shown as lines in red, are in accordance with the report of the String database.

5 Conclusion

Multi-omics data can be used in modeling gene regulatory networks. The recent rapid advances of high-throughput omics technologies have triggered the integrative multi-omics study for the in-depth understanding of the complex biological processes. However, only a few studies have considered the multi-omics data in gene regulatory network inference.

In this paper, we proposed an integrative gene regulatory network inference method, where multi-omics data and their interaction effects are integrated in the mathematical graph model. Our proposed method, iGRN, can infer gene regulatory networks from multi-omics data of CNVs and DNA methylations as well as gene expression data, and produce the homogeneous network where nodes are only genes. It enables one to analyse the gene regulatory network with most network analysis and visualisation tools efficiently. The inference capability of iGRN was assessed by the intensive experiments with simulation data. iGRN was applied to human brain data of psychiatric disorders, and the biological network of psychiatric disorders was analysed.

Acknowledgments

This research was supported in part by the National Institutes of Health/National Cancer Institute Cancer Center Support Grant (Grant number P30 CA008748).

References

- Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*. 2010; 4:132. [PubMed: 20920161]
- Aure M, et al. Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. *Genome Biology*. 2013; 14(11):R126. [PubMed: 24257477]
- Breckpot J, et al. Differences in copy number variation between discordant monozygotic twins as a model for exploring chromosomal mosaicism in congenital heart defects. *Molecular Syndromology*. 2012; 2(2):81–87. [PubMed: 22511896]
- Cai X, Bazerque J, Giannakis G. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLOS Computational Biology*. 2013; 9(5):e1003068. [PubMed: 23717196]
- Chai L, et al. A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*. 2014; 48(1):55–65. [PubMed: 24637147]
- Cisek K, Krochmal M, Klein J, Mischak H. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrology Dialysis Transplantation*. 2015; 31(12):2003–2011.
- Gustafsson M, Hörnquist M, Lombardi A. Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2005; 2(3):254–261. [PubMed: 17044188]
- Henrichsen C, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. *Human Molecular Genetics*. 2009; 18(R1):R1–R8. 238 N. Zarayeneh, et al. [PubMed: 19297395]
- Higdon R, et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *OMICS: A Journal of Integrative Biology*. 2015; 19(4):197–208. [PubMed: 25831060]
- Hill S, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*. 2016; 13(4):310–318. [PubMed: 26901648]
- Husmeier D, Werhli A. Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with Bayesian networks. *Computational Systems Bioinformatics Conference*. 2007; 6:85–95.
- Kang M, Kim D, Liu C, Gao J. Multiblock discriminant analysis for integrative genomic study. *BioMed Research International*. 2015; 10
- Kang M, Park J, Kim D, Biswas A, Liu C, Gao J. Multi-block bipartite graph for integrative genomic analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016; 99(1)
- Kim, D., et al. Integration of DNA methylation, copy number variation, and gene expression for gene regulatory network inference and application to psychiatric disorders. *Proceedings of IEEE International Conference on Bioinformatics and Bioengineering*; 2014. p. 238-242.
- Kim, D., Kang, M., Biswas, A., Liu, C., Gao, J. Integrative approach for inference of gene regulatory networks using lasso-based random featuring and application to psychiatric disorders. *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*; 2015. p. 145-150.
- Ko, E., Kang, M., Chang, H., Kim, D. Graph-theory based simplification techniques for efficient biological network analysis. *Proceedings of IEEE International Workshop on Big Data Security and Services in conjunction with IEEE Big Data Service*; 2017.
- Kristensen V, et al. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*. 2014; 5:299–313.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008; 9:559. [PubMed: 19114008]
- Liang K, Wang X. Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*. 2008; 1:253894.

- Liu C, et al. Whole-genome association mapping of gene expression in the human prefrontal cortex. *Molecular Psychiatry*. 2010; 15(8):779–784. [PubMed: 20351726]
- Njah H, Jamoussi S. Weighted ensemble learning of Bayesian network for gene regulatory networks. *Neurocomputing*. 2015; 150(Part B):404–4116.
- Oh JH, Deasy J. Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm. *BMC Bioinformatics*. 2014; 15(Suppl 7):S5.
- Omrnian N, Eloundou-Mbebi J, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific Reports*. 2016; 6(20):533.
- Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. *Bioinformatics*. 2015; 31(12):197–205.
- Rebollar E, et al. Using “omics” and integrated multi-omics approaches to guide probiotic selection to mitigate chytridiomycosis and other emerging infectious diseases. *Frontiers in Microbiology*. 2016; 7:68–86. [PubMed: 26870025]
- Reshef D, et al. Detecting novel associations in large data sets. *Science*. 2011; 334(6062):1518–1524. [PubMed: 22174245]
- Szklarczyk D, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*. 2015; 43(D1):D447–D452. [PubMed: 25352553]
- Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Statist*. 1996; 58(D1):267–288. Integration of multi-omics data for iGRN inference 239.
- Wagner J, et al. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*. 2014; 15(2):R37. [PubMed: 24555846]
- Werhli A, Husmeier D. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*. 2007; 6(1)
- Young W, Raftery A, Yeung K. Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Systems Biology*. 2014; 8(1):47. [PubMed: 24742092]
- Zhang X, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*. 2012; 28(1):98–104. [PubMed: 22088843]
- Zhang W, Li F, Nie L. Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. *Microbiology*. 2010; 156(2):287–301. [PubMed: 19910409]

Biographies

Neda Zarayeneh is pursuing her master degree in Computer Science at Texas A&M University-Commerce. She received the MSc degree in Applied Mathematics from Tehran University, Iran in 2010. Her research interests include bioinformatics, machine learning, and big data analytics.

Euiseong Ko is pursuing his master degree in Computer Science at Kennesaw State University. He received the BS degree in Electronic Engineering and Computer Science from the Hanyang University, Korea (2014). His research interests include security and privacy, big data analytics, and algorithm design.

Junghun Oh is an assistant attending in the Department of Medical Physics at Memorial Sloan Kettering Cancer Center. He develops cutting-edge computational and statistical methods, informed by bioinformatics and machine learning techniques, to identify novel diagnostic biomarkers and to build models that predict radiation treatment outcomes. He received his PhD in the Department of Computer Science from the University of Texas at Arlington and completed postdoctoral fellowships in the Department of Radiation Oncology

at Washington University School of Medicine in St. Louis and in the Department of Medical Physics at Memorial Sloan Kettering.

Sang Suh is a professor in the Department of Computer Science at Texas A&M University-Commerce. He received a M.S. in Computer Science from the University of Hawaii and a PhD in Computer Science from SMU. His research spans in the areas of data analytics, visualisation, and data mining.

Chunyu Liu is an associate professor in the Department of Psychiatry at the University of Illinois at Chicago. He received his PhD in Medical Genetics at Hunan Medical University. His research interests are at psychiatric genetics, genomics, and epigenomics.

Jean Gao received the BS degree in biomedical engineering from the Shanghai Medical University, Shanghai, China, in 1990, the MS degree in biomedical engineering from the Rose-Hulman Institute of Technology, Terre Haute, IN, in 1996, and the PhD degree in electrical engineering from Purdue University, West Lafayette, IN, in 2002. She is currently a professor with the Computer Science and Engineering Department, University of Texas at Arlington. Her research interests include medical imaging, and applications in computational biology, and clinical medical informatics, and computer vision. She is the recipient of the prestigious CAREER award from the National Science Foundation and the Outstanding Young Faculty Award from University of Texas at Arlington.

Donghyun Kim received the BS degree in Electronic and Computer Engineering from the Hanyang University, Ansan, Korea (2003), and the MS degree in Computer Science and Engineering from Hanyang University, Korea (2005). He received the PhD degree in Computer Science from the University of Texas at Dallas, Richardson, USA (2010). Currently, he is an assistant professor in the Department of Computer Science at Kennesaw State University, Marietta, USA. From 2010 to 2016, he was an assistant professor in the Department of Mathematics and Physics at North Carolina Central University, Durham, USA. His research interests include security and privacy, social computing, mobile computing, cyber physical systems, wireless and sensor networking, and algorithm design and analysis. He is an associate editor of Discrete Mathematics, Algorithms and Applications. He is a member of ACM and a senior member of IEEE.

Mingon Kang is an assistant professor in the Department of Computer Science at Kennesaw State University. He received his master's and PhD degrees in Computer Science at the University of Texas at Arlington. His research interests include bioinformatics, machine learning, and big data analytics.

Figure 1.

A simple integrative gene regulatory network. The interaction effects of copy number variation (CNV) and DNA methylation (DM) are incorporated in the gene regulatory network model

Figure 2.
Overall ROC curves

Figure 3.
TPRs for interaction effects of CNVs and DNA methylations

Figure 4.
Sensitivity

Figure 5.
Sensitivity on copy number variations and DNA methylations

Figure 6.
False discovery rate

Figure 7.
Volcano plot

Figure 8.
The integrative gene regulatory network inferred by iGRN

Figure 9. The sub-components of the integrative gene regulatory network that maximise weights of the components

Table 1

AUROC with simulation data

Methods	GRN	DCGRN	iGRN
AUROC	0.895	0.843	0.938