Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 8520480, 10 pages
https://doi.org/10.1155/2017/8520480

Research Article

Node-Structured Integrative Gaussian Graphical Model Guided by Pathway Information

SungHwan Kim,[1,2] Jae-Hwan Jhong,[3] JungJun Lee,[3] Ja-Yong Koo,[3] ByungYong Lee,[4] and SungWon Han[5]

[1]*Department of Statistics, Keimyung University, Daegu, Republic of Korea*
[2]*The Institute of Natural Science, Keimyung University, Daegu, Republic of Korea*
[3]*Department of Statistics, Korea University, Seoul, Republic of Korea*
[4]*Graduate School of Information Security, Korea University, Seoul, Republic of Korea*
[5]*School of Industrial Management Engineering, Korea University, Seoul, Republic of Korea*

Correspondence should be addressed to ByungYong Lee; rom109101@gmail.com and SungWon Han; swhan@korea.ac.kr

Received 31 October 2016; Revised 20 February 2017; Accepted 6 March 2017; Published 12 April 2017

Academic Editor: Hongmei Zhang

Up to date, many biological pathways related to cancer have been extensively applied thanks to outputs of burgeoning biomedical research. This leads to a new technical challenge of exploring and validating biological pathways that can characterize transcriptomic mechanisms across different disease subtypes. In pursuit of accommodating multiple studies, the joint Gaussian graphical model was previously proposed to incorporate nonzero edge effects. However, this model is inevitably dependent on post hoc analysis in order to confirm biological significance. To circumvent this drawback, we attempt not only to combine transcriptomic data but also to embed pathway information, well-ascertained biological evidence as such, into the model. To this end, we propose a novel statistical framework for fitting joint Gaussian graphical model simultaneously with informative pathways consistently expressed across multiple studies. In theory, structured nodes can be prespecified with multiple genes. The optimization rule employs the structured input-output lasso model, in order to estimate a sparse precision matrix constructed by simultaneous effects of multiple studies and structured nodes. With an application to breast cancer data sets, we found that the proposed model is superior in efficiently capturing structures of biological evidence (e.g., pathways). An R software package nsiGGM is publicly available at author's webpage.

1. Introduction

Genomic data have been extensively applied to analyze disease mechanism on the basis of predictive signatures from DNA alterations (e.g., genotyping and mutation), RNA transcription (e.g., gene or isoform expression and fusion transcripts), and gene regulation by epigenetic changes (e.g., methylation, protein-DNA interaction, and miRNA expression). In particular, gene regulation is a complicated system that builds on tens of thousands of cellular components' interactions and diverse activities across multiple layers. Biological networks are the most popularly used data resource to sketch this interconnectivity of gene regulations. High-throughput genomic technologies are paving the way toward systematically characterizing diverse types of biological networks and suggestive of underlying gene regulation mechanisms. And yet a complete inference of network's complexity has been a long concern in the field of systems biology.

To circumvent the shortcoming of single feature-based analysis, the activity of a gene or of a whole biological process in a disease can be assessed by sets of genes (a.k.a. gene set enrichment analysis or pathway analysis). In doing so, a bulk of pathways have been identified through many cancer-related researches [1]. Pathway information demonstrates cellular functions and biological processes or represents a unique signature of deregulation of a given gene [2]. For

example, the pathway or signature associated with the activity of a given oncogene is defined as the set composed of those genes most differentially expressed by perturbation of oncogenes [3–5]. Importantly, the usage of pathway information is increasingly prevalent in biomedicine. For instance, target drug associated with potential pathway is taken as a practical solution to overcome the traditional drug discovery that usually adopts the one-drug-one-target approach. This strategy takes into account the fact that the disease occurrence is usually the result of complex interactions of molecular events.

In recent years, large-scale genomic data generated from relevant biological experiments or clinical hypotheses have increasingly soared, as high-throughput experiment technologies have markedly advanced [6]. Such increasing genomic data has been publicly available in data repositories (e.g., Gene Expression Omnibus and Sequence Read Archive). This abundance of biological experiments poses a new challenge of multiple data in regard to exploring and validating biological signatures and pathways. More precisely, a question of network analysis often relates to how to characterize underlying transcriptomic patterns or molecular mechanisms across disease subtypes or between case-control groups, because it is commonplace that biological signals are not coherently present across studies. Generally a single network [7–9] is found to accurately estimate underlying dependency with an adjustment of gene perturbation effects (e.g., polymorphic genotype alteration [10, 11]). Nonetheless, these methods hardly discover network patterns of subtle signals and dynamic features in the midst of coupled networks under diverse conditions. Moreover, single networks potentially generate many potential false positive signals (edges) attributed to experimental biases and errors. To address this challenge, the recent trend of data analysis has been in the spotlight to data integration allowing for multiple data to achieve a more accurate network inference. To this end, many have proposed methods to combine multiple networks based on unified model [12–14]. This approach is also known as integrative analysis and is analogue to traditional meta-analysis.

The joint Gaussian graphical model (JGGM; Danaher et al. [12]) focuses on incorporating nonzero edge effects (i.e., off-diagonal entries of precision matrix) to combine multiple studies in view of integrative analysis. This model, however, inevitably is dependent on post hoc analysis when validating biological significance. Therefore it is interesting to combine not only DNA and/or transcriptomic changes but also pathway information as such well-ascertained biological evidence. Normally we perform post hoc analysis to see if the estimated gene networks are enriched for any pathways. Contrary to this, it is also sensible to estimate gene networks, with an adjustment of pathway information. It is common that we hardly combine pathway information in spite of its biological significance. To the best of our knowledge, no method has been proposed that can accommodate overlapping node structures, mainly due to overlapped gene annotations of pathway gene sets. To tackle this problem, we propose a new graphical model called "node-structured integrative Gaussian graphical model (nsiGGM)" jointly leveraging a priori knowledge of pathway information. This method allows for overlapping group lasso problems, making it possible to integrate overlapped genes of pathways. It is worthwhile for biological pathways to intervene the network estimation to reveal true gene regulatory network. The nsiGGM builds on prespecified structured nodes with multiple genes as building blocks in the stage of estimating a precision matrix. The implementation rule employs l_1/l_2 lasso penalty of structured input-output lasso model [15], in order to estimate sparse precision matrix that accounts for simultaneous effects of multiple studies and structured nodes. With an application to simulated and breast cancer genomic data, the proposed model is found to be superior in efficiently capturing transcriptional modules predefined by pathway database. A software package (nsiGGM) is publicly available at author's webpage (https://sites.google.com/site/sunghwanshome/).

This paper is outlined as follows. In Section 2, we review background knowledge of the standard and joint Gaussian graphical models. In addition, we propose the node-structured integrative Gaussian graphical model (nsiGGM). In Section 3, we describe an implementation strategy that is primarily based on the input-output lasso. In Section 4, we compare performance of our proposed methods with other methods using real breast cancer data (TCGA) and simulated data. In Section 5, conclusions and further studies are discussed.

2. Method

In this section, we briefly discuss methodological backgrounds on the Gaussian graphical models (GGM) aiming at constructing gene networks. In what follows, we propose the node-structured integrative Gaussian graphical model (nsiGGM) that can accommodate a priori biological knowledge (e.g., pathway data or targeted predictive genes of miRNA).

2.1. Gaussian Graphical Models for Gene Networks. A Gaussian graphical model demonstrates the conditional dependency of multiple random variables, $Y_1, \ldots, Y_p$, with a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is a set of nodes and E is a set of edges indicating that nodes are linked and conditionally dependent. Let Y follow the multivariate Gaussian distribution $N_p(0, \Sigma)$, where Σ is a $p \times p$ covariance matrix. Let $\Sigma^{-1} = \Theta$ denote the inverse covariance matrix (also known as a precision matrix). More precisely, each nonzero off-diagonal element θ_{ij} implies conditional dependency between the ith and jth nodes given all the other variables, $i, j = 1, \ldots, p$, whereas the covariance Σ presents marginal dependencies without considering other variables. This model is also called a GGM [16]. The graphical lasso [9, 17] produces a sparse Gaussian graphical model constructed in nonpenalized edges in Θ. The graphical lasso minimizes the negative log-likelihood with the L_1 lasso penalty:

$$\arg\min_{\Theta} - \log\det\Theta + \operatorname{tr}(S\Theta) + \lambda \|\Theta\|_1, \tag{1}$$

where $\operatorname{tr}(A)$ is the trace of matrix A, S is the sample covariance matrix, and $\|\Theta\|_1 = \sum_i \sum_j |\theta_{ij}|$ is the regularization parameter

adjusting the degree of sparsity. The optimal value for λ can be chosen by cross-validation or the Bayesian information criterion (BIC; Schwarz [18]; Yuan and Lin [8]).

2.2. Joint Gaussian Graphical Models for Combining Multiple Studies. In this section, we revisit the joint Gaussian graphical models (JGGM) proposed by Danaher et al. [12]. Simply put, the JGGM combines multiple studies and constructs multiple networks in a unified model. Let K denote the number of studies in our data and $\{\Sigma^{-1}\} = (\Sigma_1^{-1}, \ldots, \Sigma_K^{-1})$ the true precision matrices. Consider genomic data of K studies, $Y^{(1)}, \ldots, Y^{(K)}$, each of which consists of n_k samples with p common features, where $K \geq 2$. We assume that $\sum_{k=1}^{K} n_k$ observations are independent and that those of each data set follow the multivariate normal distribution as $Y^{(k)} \sim N_p(\mu_k, \Sigma_k)$ for $1 \leq k \leq K$. It is well known in meta-analysis that multiple data sets are of common associations and genomic characteristics among features (e.g., genetic association intensity). It, therefore, is worth estimating precision matrices across K studies in parallel rather than separate estimation. To this end, we assume that the features within each data set are centered and take the form of a penalized log-likelihood with the group sparsity-inducing L_2 penalty that maximizes (2) with respect to $\{\Theta\}$:

$$\sum_{k=1}^{K} n_k \left[\log\left\{\det\left(\Theta^{(k)}\right)\right\} - \operatorname{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] + \lambda_1 \sum_{k=1}^{K}\sum_{i\neq j}\left|\theta_{ij}^{(k)}\right| + \lambda_2 \sum_{i\neq j}\left(\sum_{k=1}^{K}\theta_{ij}^{(k)2}\right)^{1/2} \tag{2}$$

subject to $\{\Theta\} = (\Theta^{(1)}, \ldots, \Theta^{(K)})$ being positive definite, where $S^{(k)} = (1/n_k)(Y^{(k)})^T Y^{(k)}$ is the sample covariance matrix of $Y^{(k)}$ and λ_1, λ_2 are nonnegative tuning parameters. It is interesting to note that the L_2-penalty captures similarity across the K precision matrices. Due to this property, the penalty terms of (2) are also referred to as the joint graphical lasso (JGL). Moreover, the L_1 penalty induces estimated precision matrices to be sparse.

2.3. Node-Structured Integrative Gaussian Graphical Model. In this section, we propose an integrative graphical model that can accommodate a priori known structure of genomic features. Learning gene networks, the sparseness of precision matrix can be guided to some extent by known feature modules (e.g., pathway information). Typically data integration allows picturing the interplay of underlying biological factors. In this regard, it is worthwhile accommodating known feature module information ascertained in previous experiments. In doing so, we seek to integrate a priori feature module information to be embedded across multiple networks via an additional L_2 group penalty. The following objective function is taken to minimize

$$\sum_{k=1}^{K} n_k \left[\log\left\{\det\left(\Theta^{(k)}\right)\right\} - \operatorname{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] + \lambda_1 \sum_{k=1}^{K}\sum_{i\neq j}\left|\theta_{ij}^{(k)}\right| + \lambda_2 \sum_{i\neq j}\left(\sum_{k=1}^{K}\theta_{ij}^{(k)2}\right)^{1/2} + \lambda_3 \sum_{k=1}^{K}\sum_{\mathbf{g}_m\in G}\left\|\Theta_{\mathbf{g}_m}^{(k)}\right\|_2, \tag{3}$$

where $\mathbf{g}_m$ is a subset of off-diagonal entry indices of Θ for $0 \leq m \leq M$, $G = \{\mathbf{g}_1, \ldots, \mathbf{g}_M\}$, M is the number of a priori feature modules, and $\|\Theta_{\mathbf{g}_m}^{(k)}\|_2 = (\sum_{(i,j)\in\mathbf{g}_m} \theta_{ij}^{(k)2})^{1/2}$. Importantly, it is noted that elements of $\mathbf{g}_m$ can be overlapped (e.g., duplicated genes of two different pathways). The third penalty, $\|\Theta_{\mathbf{g}_m}^{(k)}\|_2$ adjusted by $\lambda_3 \geq 0$ pertains to structured feature modules (i.e., structured node in networks) on the basis of a priori known information. Here, unbiased regularization to each feature should be taken into consideration, in the sense that the feature overlapping inevitably comes into play.

In what follows, we present a toy example to demonstrate how a priori information constructs feature modules in $\Theta_{\mathbf{g}_m}^{(k)}$. In Figure S1, in Supplementary Material available online at https://doi.org/10.1155/2017/8520480, we take an example of networks consisting of 5 common nodes (e.g., genomic features) across three studies. In Figure S1A, the second penalty with λ_2 captures matched up common edges (e.g., $\theta_{14}^{(1)}, \theta_{14}^{(2)}, \theta_{14}^{(3)}$) identical to the joint graphical lasso. Besides, the third group lasso penalty with λ_3 accommodates the six edges of the three features in a predefined module $\Theta_{\mathbf{g}_1}^{(k)}$ so that feature regulatory effects can be further modeled in the context of data integration (see Figure S1B). Importantly note that this module structure (e.g., pathway) is priorly known knowledge. It is interesting that this approach is in line with the integrative cluster [19] that allows for *cis*-regulatory effects and target gene prediction for miRNAs. In the case of multiple modules in network, suppose that we are given a set of five genes $\{X_i\}$ and a precision matrix $\{\theta_{ij}\}$ for $1 \leq i, j \leq 5$. Let a priori information generate two feature modules defined as Module 1, $\{X_1, X_2, X_3, X_4\}$, and Module 2, $\{X_3, X_4, X_5\}$, and then we can enumerate precision matrix's index (i, j) of each module for all i, j, say, $\mathbf{g}_1 = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}$ and $\mathbf{g}_2 = \{(3,4), (3,5), (4,5)\}$. Of note, the component $(3,4)$ is simultaneously present in both $\mathbf{g}_1$ and $\mathbf{g}_2$, implicating that a suitable implementation is required for regularization to the overlapped component $(3,4)$. To estimate solutions to (3), we apply the structured input-output lasso [15] that can handle overlapped features, making it possible to learn a model allowing for both single-node effects across studies and predefined node structures (e.g., pathway modules). Inspired by integrative nature of this method, we call this graphical model the node-structured integrative Gaussian

graphical model (nsiGGM). When it comes to tuning the penalty parameters (λ_1, λ_2, and λ_3), the BIC is applied to determine the optimal sparseness of networks' edges.

3. Implementation Strategy

3.1. Structured Alternating Directions Method of Multipliers Algorithm. In this section, we delineate the implementation strategy for the nsiGGM. We solve problem (3) by using structured alternating directions method of multipliers algorithm (sADMM). The alternating directions method of multipliers algorithm (ADMM) was previously introduced to tackle the problem of the JGL [12]. Similar to the JGL, the sADMM proposed in spirit of the ADMM is designed to adopt the structured input-output lasso in order to embed node structures into the model. We first reformulate (3) with $P(\Theta)$ and Z as

$$\underset{\Theta,Z}{\text{maximize}} \quad \left(-\sum_{k=1}^{K} n_k \left[\log\left\{ \det\left(\Theta^{(k)}\right)\right\} - \operatorname{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] + P\left(\{Z\}\right)\right), \tag{4}$$

where $P(\{Z\}) = \lambda_1 \sum_{k=1}^{K}\sum_{i\neq j} |Z_{ij}^{(k)}| + \lambda_2 \sum_{i\neq j}(\sum_{k=1}^{K} Z_{ij}^{(k)2})^{1/2} + \lambda_3 \sum_{k=1}^{K}\sum_{\mathbf{g}_m\in G} \|Z_{\mathbf{g}_m}^{(k)}\|_2$; $Z^{(k)} = \Theta^{(k)}$ for $1 \le k \le K$ and $\{Z\} = (Z^{(1)}, \ldots, Z^{(K)})$ that satisfies positive definiteness. Boyd et al. [20] proposed the scaled augmented Lagrangian to solve problem (4) by

$$\begin{aligned} &L\left(\{\Theta\}, \{Z\}, \{U\}\right) \\ &= -\sum_{k=1}^{K} n_k \left[\log\left\{ \det\left(\Theta^{(k)}\right)\right\} - \operatorname{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] \\ &\quad + P\left(\{Z\}\right) + \frac{1}{2}\sum_{k=1}^{K} \left\|\Theta^{(k)} - Z^{(k)} + U^{(k)}\right\|_F^2 \\ &\quad - \frac{1}{2}\sum_{k=1}^{K} \left\|U^{(k)}\right\|_F^2, \end{aligned} \tag{5}$$

where $\{U\} = (U^{(1)}, \ldots, U^{(K)})$ are dual variables and $\|A\|_F$ denotes the Frobenius norm of matrix A (i.e., $\|A\|_F = \sqrt{\sum_i \sum_j A_{ij}^2}$). The sADMM algorithm repeatedly solves the three-step optimization with respect to $\Theta_{(i)}$, $Z_{(i)}$, and $U_{(i)}$, starting with initial values of the related parameters: $\Theta^{(k)} = I$, $U^{(k)} = 0$, and $Z^{(k)} = 0$ for $1 \le k \le K$. The iteration is repeated until convergence as follows: In Θ-step for $1 \le k \le K$, update $\Theta^{(k)}$ that minimizes

$$\begin{aligned} &-\sum_{k=1}^{K} n_k \left[\log\left\{ \det\left(\Theta^{(k)}\right)\right\} - \operatorname{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] \\ &\quad + \frac{1}{2}\sum_{k=1}^{K} \left\|\Theta^{(k)} - Z_{(i-1)}^{(k)} + U_{(i-1)}^{(k)}\right\|_F^2. \end{aligned} \tag{6}$$

In Z-step, for $k = 1, \ldots, K$, update $Z^{(k)}$ that minimizes

$$\begin{aligned} &\frac{1}{2}\sum_{k=1}^{K} \left\|Z^{(k)} - A_{(i)}^{(k)}\right\|_F^2 + \lambda_1 \sum_{k=1}^{K}\sum_{i\neq j} \left|Z_{ij}^{(k)}\right| \\ &\quad + \lambda_2 \sum_{i\neq j} \left(\sum_{k=1}^{K} Z_{ij}^{(k)2} \right)^{1/2} + \lambda_3 \sum_{k=1}^{K}\sum_{\mathbf{g}_m\in G} \left\|Z_{\mathbf{g}_m}^{(k)}\right\|_2, \end{aligned} \tag{7}$$

where $A_{(i)}^{(k)} = \Theta_{(i)}^{(k)} + U_{(i-1)}^{(k)}$. To find the optimal solution of (7), we directly apply the structured input-output lasso [15] to (7) using both coordinate descent algorithm and KKT conditions considered to boost up the computational speed. For more details, see [15]. In U-step, for $k = 1, \ldots, K$, update $U^{(k)}$ as $U_{(i-1)}^{(k)} + \Theta_{(i)}^{(k)} - Z_{(i)}^{(k)}$. Update repeatedly the three parameters until convergence by a stopping rule below:

$$\frac{\sum_k \left\|\Theta_{(i)}^{(k)} - \Theta_{(i-1)}^{(k)}\right\|_1}{\sum_k \left\|\Theta_{(i-1)}^{(k)}\right\|_1} < 10^{-3}. \tag{8}$$

Putting together, Algorithm 1 encapsulates the structured alternating directions method of multipliers algorithm.

4. Numerical Studies

4.1. Simulated Data. In this section, we carry out experimental studies to assess performance of the nsiGGM. In brief, the following describes how we generate simulated data. The experimental scheme is largely motivated by Chun et al. [14]. Let K be the total number of studies, each containing true signal genes $p = 40$ for a priori module (e.g., pathway genes) and sample size $n^{(k)} = 100$, where $1 \le k \le K$ (=3). Starting off with edges of signal genes, we first generate network edges of 100 nodes subject to the scale-free network structures, the most commonly observed structures in biology, being simulated by applying the Barabasi Albert algorithm [21]. Subsequent to this, we randomly added four edges to impose random effects. Constructing network structures, we simulate the precision matrices by setting values of the off-diagonals sampled from $\text{Unif}(-0.1, 0.1)$ and by setting the diagonal elements with $\sum_{j\neq i} |\theta_{i,j}^{(k)}|$. The process is repeated until $\Theta^{(k)}$ becomes a positive definite matrix. For simulating $Y^{(k)}$, we first consider a scenario such that no covariate incurs dependency among genes. Thus, this is an ideal experiment scenario in that any conditional dependency is not taken into consideration to the model. We simulated $Y^{(k)}$, where each ith row of $Y^{(k)}$ was randomly sampled from $N_p(0, \Theta^{(k)^{-1}})$. Simulations were repeated and average values are presented in Tables 1 and 2. To examine performance of the nsiGGM,

(1) Initialize $\Theta^{(k)} = I, U^{(k)} = 0$ and $Z^{(k)} = 0$ for $k = 1, \ldots, K$.
(2) Update $\Theta^{(k)}$, $Z^{(k)}$ and $U^{(k)}$ until convergence for $k = 1, \ldots, K$:
(i) Θ-step
Update $\Theta^{(k)}$ that minimizes
$$-\sum_{k=1}^{K} n_k \left[\log\left\{\det\left(\Theta^{(k)}\right)\right\} - \operatorname{tr}\left(S^{(k)}\Theta^{(k)}\right)\right] + \frac{1}{2}\sum_{k=1}^{K}\left\|\Theta^{(k)} - Z_{(i-1)}^{(k)} + U_{(i-1)}^{(k)}\right\|_F^2$$
(ii) Z-step
For $k = 1, \ldots, K$, update $Z^{(k)}$ that minimizes
$$\frac{1}{2}\sum_{k=1}^{K}\left\|Z^{(k)} - A_{(i)}^{(k)}\right\|_F^2 + \lambda_1\sum_{k=1}^{K}\sum_{i\neq j}\left|Z_{ij}^{(k)}\right| + \lambda_2\sum_{i\neq j}\left(\sum_{k=1}^{K} Z_{ij}^{(k)^2}\right)^{1/2} + \lambda_3\sum_{k=1}^{K}\sum_{\mathbf{g}_m\in G}\left\|Z_{\mathbf{g}_m}^{(k)}\right\|_2,$$
where $A_{(i)}^{(k)} = \Theta_{(i)}^{(k)} + U_{(i-1)}^{(k)}$.
(iii) U-step
For $k = 1, \ldots, K$, update $U^{(k)}$ as $U_{(i-1)}^{(k)} + \Theta_{(i)}^{(k)} - Z_{(i)}^{(k)}$.
(3) Output $\widehat{\Theta} = (\widehat{\Theta}^{(1)}, \widehat{\Theta}^{(2)}, \ldots, \widehat{\Theta}^{(K)})$.

Algorithm 1: The structured alternating directions method of multipliers algorithm.

Table 1: Performance comparisons of the nsiGGM with the JGGM and GGM using data simulated along with predefined module genes.

Methods	# of noise genes	Sensitivity (s.e.)	Specificity (s.e.)	Youden (s.e.)
nsiGGM	30	0.2217 (0.0253)	0.9433 (0.0036)	0.1650 (0.0257)
	40	0.2125 (0.0133)	0.9472 (0.0053)	0.1598 (0.0117)
	50	0.2034 (0.019)	0.9481 (0.0035)	0.1515 (0.0175)
JGGM	30	0.2433 (0.04)	0.8685 (0.0273)	0.1118 (0.0161)
	40	0.2815 (0.0418)	0.8321 (0.0309)	0.1136 (0.0146)
	50	0.1920 (0.0425)	0.8733 (0.0318)	0.0653 (0.0124)
GGM	30	0.2593 (0.0264)	0.8325 (0.0214)	0.0918 (0.0094)
	40	0.2752 (0.029)	0.8050 (0.0257)	0.0802 (0.0074)
	50	0.2177 (0.0303)	0.8431 (0.0268)	0.0608 (0.0085)

Table 2: Shown are the brief descriptions of the three data information pieces used in real genomic application.

Study	Data type	# of samples	# of matched genes	Reference
Breast cancer	mRNA	319	10,676	The Cancer Genome Atlas (TCGA)
Breast cancer	mRNA	134	10,676	GSE7390
Breast cancer	mRNA	209	10,676	GSE2034

sensitivity, specificity, and Youden index were benchmarked by comparing the JGGM [12] and GGM [16]. Youden index is defined as Sensitivity + Specificity − 1, ranging from −1 to 1. In principle, the higher the Youden index, the higher the prediction accuracy.

In Table 1, Youden index of the nsiGGM appears to be clearly declining as noise edges increase in number and yet is consistently larger than that shown in the JGGM and GGM. This is mainly due to the fact that the JGGM suffers low specificity (0.8685–0.8733) compared to the nsiGGM (0.9433–0.9481). In contrast, the JGGM slightly outperforms, when 30 and 40 noises are augmented, the nsiGGM for sensitivity at the expense of poor specificity. Taken together, it is clear to say that the nsiGGM is superior to the JGGM and GGM in detecting the true underlying pathway sets.

4.2. Application to Genomic Data. In this section, we demonstrate applications to three mRNA expression profiles for breast cancer. We collected two microarray profiles from Desmedt et al. [22], Wang et al. [23], and TCGA cancers data from TCGA's web portal (https://cancergenome.nih.gov/), where we retrieved mRNA data of breast carcinoma (BRCA). We matched up features across all studies and filtered out probes by the rank sum of mean and standard deviation

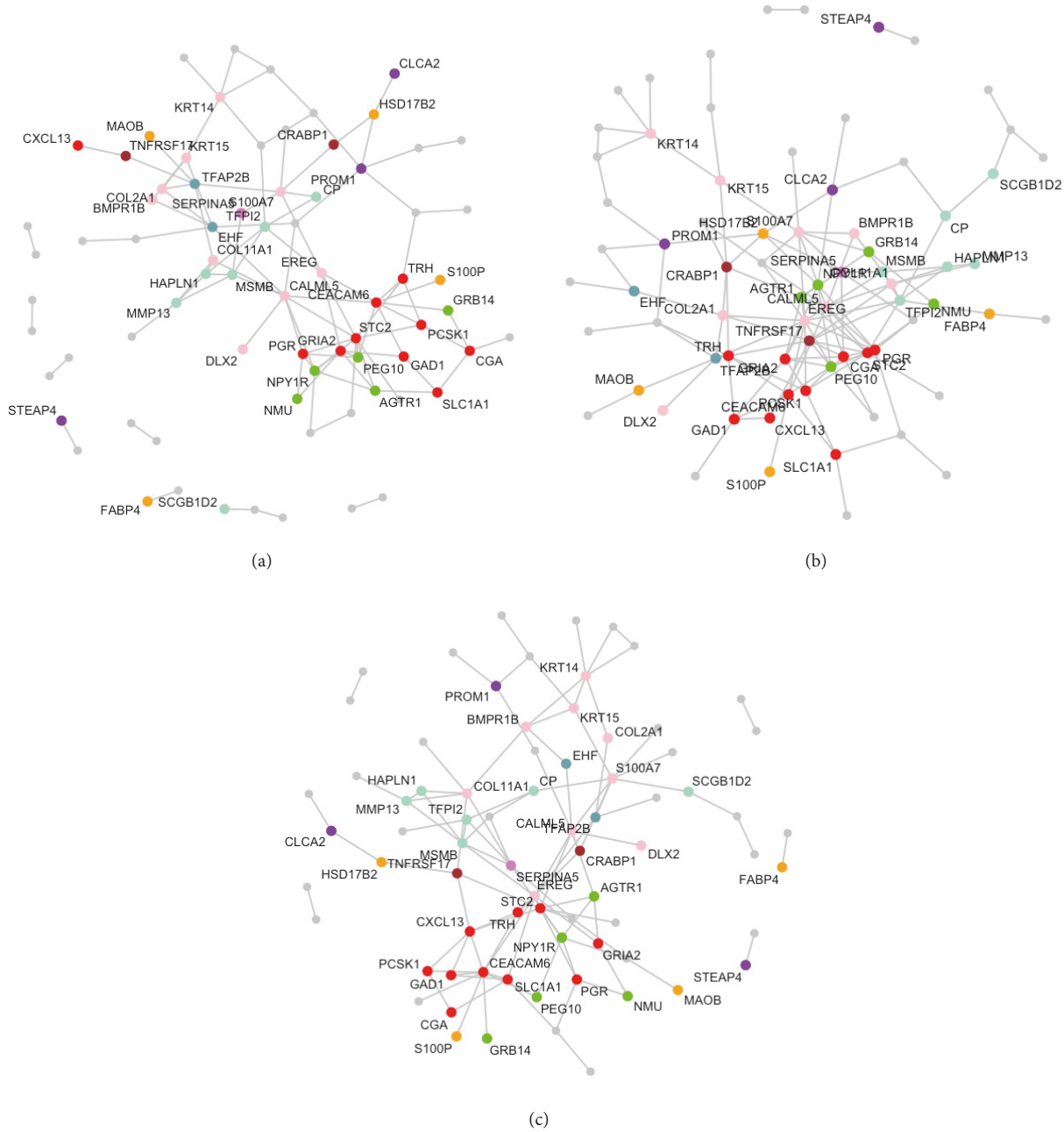

FIGURE 1: Three gene networks estimated by the nsiGGM. The detection rate of pathway genes is 0.573.

(SD < 0.99; Wang et al. [24]), which selected 106 genes. Table 2 delineates detailed information of miRNA expression data. In what follows, we examine if the nsiGGM is suited to improve accuracy for detecting pathway genes. We collected gene sets from exploring the Molecular Signatures Database (MSigDB) v2.5 gene set collections [25], consisting of at least 11 genes of 106 genes, of which 53 distinct genes belong to the 11 pathways presented in Table 3. To evaluate a detection rate of pathway genes, we define an evaluation benchmark, $R(\cdot)$ as follows:

$$R\left(\Phi_j\right) = \frac{\sum_{i\in\Phi_j} I\left(i\text{th gene of } j\text{th network's node belongs to any of given pathways}\right)}{\#\text{ of total selected nodes}}, \tag{9}$$

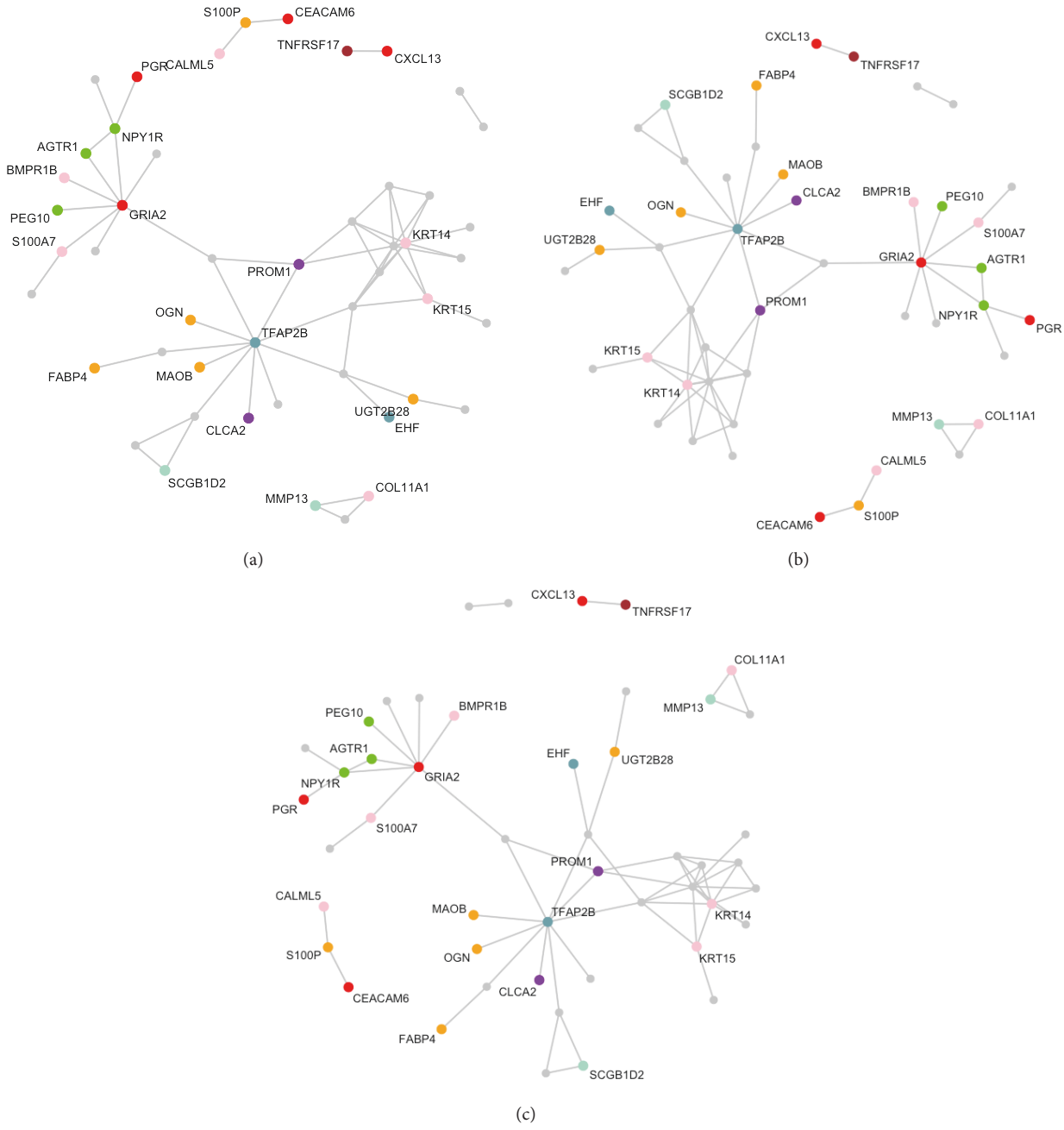

Figure 2: Three gene networks estimated by the JGGM. The detection rate of pathway genes is 0.521.

where Φ_j is a set of gene indices, whose genes form jth network and $I(\cdot)$ is an indicator function. Comparing the JGGM, we examine whether the nsiGGM effectively captures the existing pathway structures better than the JGGM in context of connectivity and proportions of identified pathway genes. To first appearances, the nsiGGM effectively represents modules well enriched with pathway genes in Figure 1, as compared to those of the JGGM in Figure 2. In support of this notion, given that we observed $\sum_j R(\Phi_j)$ of nsiGGM = 0.573 and $\sum_j R(\Phi_j)$ of JGGM = 0.521, where $1 \leq j \leq 3$, it is not surprising to say that the nsiGGM can facilitate constructing gene networks biologically more enriched for pathway gene sets than the JGGM. Table 3 enumerates the pathway genes discovered by the nsiGGM, each being highlighted by bold and underlined characters (note: asterisks represent pathway genes identified by only the nsiGGM not by JGGM). Interestingly, there are many pathways genes monitored by the nsiGGM, but not by the JGGM. Focusing on the cell signaling pathway, we particularly notice that EREG [26], SLC1A1 [27], STC2 [28], GAD1 [29], and TRH [30] are genes not selected

Table 3: The pathway sets from the Molecular Signatures Database (MSigDB) analyzed in the nsiGGM. (Note: asterisks represent pathway genes identified by only the nsiGGM not by JGGM.)

Pathway 1: extracellular region (11 genes)
SERPINA5*, **MMP13**, **EREG***, **HAPLN1***, **CP***, **S100A7**, CRISP3, **SCGB1D2**, **COL11A1**, **TFPI2***, **MSMB***
Pathway 2: membrane part (11 genes)
EREG*, PTPRN2, **CLCA2**, **NPY1R**, TRPA1, **TNFRSF17**, **AGTR1**, **CEACAM6**, **SLC1A1***, **PROM1**, **HSD17B2***
Pathway 3: membrane (14 genes)
EREG*, PTPRN2, **STEAP4***, **GRIA2**, **CLCA2**, **NPY1R**, TRPA1, **TNFRSF17**, **AGTR1**, **SERPINA5***, **CEACAM6**, **SLC1A1***, **PROM1**, **HSD17B2***
Pathway 4: cytoplasm (13 genes)
OGN, CA2, MYBPC1, NLRP2, **MAOB**, **UGT2B28**, **S100A7**, CRISP3, **PEG10**, **S100P**, **FABP4**, CLGN, **HSD17B2***
Pathway 5: plasma membrane (12 genes)
AGTR1, **CEACAM6**, **EREG***, PTPRN2, **SLC1A1***, **STEAP4***, **PROM1**, **GRIA2**, **CLCA2**, **NPY1R**, TRPA1, **TNFRSF17**
Pathway 6: system development (12 genes)
EREG*, **TFAP2B**, **CALML5**, **KRT15**, **KRT14**, **DLX2***, **BMPR1B**, MSTN, **S100A7**, NKX2-2, **COL11A1**, **COL2A1***
Pathway 7: signal transduction (15 genes)
EREG*, **CALML5**, **GRIA2**, **TRH***, **PGR**, **NPY1R**, **CGA***, **CRABP1***, **TNFRSF17**, **AGTR1**, **CEACAM6**, **PEG10**, **GRB14***, **STC2***, **NMU***
Pathway 8: multicellular organismal development (15 genes)
EREG*, **TFAP2B**, **CALML5**, **KRT15**, **KRT14**, DLX2, **BMPR1B**, MSTN, **S100A7**, NKX2-2, **COL11A1**, **EHF**, **COL2A1***, **CRABP1***, **TNFRSF17**
Pathway 9: cell signaling (11 genes)
PGR, **CEACAM6**, **EREG***, **PCSK1***, **CXCL13**, **SLC1A1***, **CGA***, **STC2***, **GAD1***, **GRIA2**, **TRH***
Pathway 10: anatomical structure development (13 genes)
EREG*, **TFAP2B**, **CALML5**, **KRT15**, **KRT14**, **DLX2***, **BMPR1B**, MSTN, **S100A7**, NKX2-2, **COL11A1***, **COL2A1**, **EHF**
Pathway 11: organ development (11 genes)
MSTN, **EREG***, **CALML5**, **S100A7**, **KRT15**, **KRT14**, NKX2-2, **DLX2***, **COL11A1**, **COL2A1***, **BMPR1B**

by the JGGM but nonetheless previously were monitored in signaling pathways. Importantly, Hou et al. [28] showed that STC2 inhibited tumorigenesis and metastasis of breast cancer cells, indicating that STC2 may inhibit epithelial-mesenchymal transition (EMT) at least partially through the PKC/Claudin-1-mediated signaling in human breast cancer cells. Therefore, STC2 can be taken into consideration as a potential biomarker for metastasis and targeted therapy in human breast cancer. Besides, signaling through glutamate receptors in regard to SLC1A1 has been reported in human cancers [31]. In support of this evidence, it is also well known that increases in SLC1A1 expression subject to hypoxia-inducible factors (HIFs) possibly contributes to increased efflux of glutamate, by which glutamate transporters and receptors are regulated to activate key signal transduction pathways that promote cancer progression [27]. Therefore, it is clear to say that the nsiGGM is superior in detecting genes capable of implicating the functional process of human cancers in essence.

5. Conclusion and Discussion

In this article, we propose a new graphical model called node-structured integrative Gaussian graphical model (nsiGGM) jointly learning Gaussian graphical models with an emphasis of prior knowledge of pathway information. It is highlighted that this method allows us to handle overlapping group lasso problems, making it possible to integrate overlapped pathway gene sets. With applications to experimental and real data, we verified outstanding numerical performance of the nsiGGM and analytical capability of inducing biological significance related to breast cancer. And yet it might be controversial whether prior knowledge too excessively determines the network structures. Despite apprehension to overly guided network structures, a priori known information can be still acceptable in that the nsiGGM selects tuning parameters on the basis of the likelihood-based BIC.

The proposed nsiGGM is highly subject to computational complexity in nature, mainly due to the coordinate decent algorithm to tackle the sparse overlapping group lasso. Since the sparse overlapping group lasso applied here deals with both study-specific effects and prior knowledge, the optimization becomes inevitably complicated. Our current package is implemented in R and the routine flows can be further expedited via C/C++ in the future. Currently, the prior knowledge of regulatory structure is accommodated to an unidirectional graphical model. It is also interesting that we impose the prior knowledge to directional networks instead, so that the presence or absence of directional edges

amid multiple features can be explicitly modeled. We leave these tasks for future research.

Disclosure

SungWon Han is the corresponding author, and ByungYong Lee is the cocorresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Acknowledgments

This research is supported by the Korea University Grant (K1607901) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2017R1C1B5017528).

References

[1] R. Maglietta, A. Distaso, A. Piepoli et al., "On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers," *Journal of Biomedical Informatics*, vol. 43, no. 3, pp. 397–406, 2010.

[2] G. A. Viswanathan, J. Seto, S. Patil, G. Nudelman, and S. C. Sealfon, "Getting started in biological pathway construction and analysis," *PLoS Computational Biology*, vol. 4, no. 2, article e16, 2008.

[3] A. H. Bild, G. Yao, J. T. Chang et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.

[4] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[5] T. Kessler, H. Hache, and C. Wierling, "Integrative analysis of cancer-related signaling pathways," *Frontiers in Physiology*, vol. 4, article 124, 2013.

[6] S. Richardson, G. C. Tseng, and W. Sun, "Statistical methods in integrative genomics," *Annual Review of Statistics and Its Application*, vol. 3, pp. 181–209, 2016.

[7] H. Li and J. Gui, "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, vol. 7, no. 2, pp. 302–317, 2006.

[8] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[10] J. Yin and H. Li, "A sparse conditional Gaussian graphical model for analysis of genetical genomics data," *Annals of Applied Statistics*, vol. 5, no. 4, pp. 2630–2650, 2011.

[11] L. Zhang and S. Kim, "Learning gene networks under SNP perturbations using eQTL datasets," *PLoS Computational Biology*, vol. 10, no. 2, Article ID e1003420, 2014.

[12] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 76, no. 2, pp. 373–397, 2014.

[13] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1–15, 2011.

[14] H. Chun, M. Chen, B. Li, and H. Zhao, "Joint conditional Gaussian graphical models with multiple sources of genomic data," *Frontiers in Genetics*, vol. 4, article 294, 2013.

[15] S. Lee and E. P. Xing, "Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs," *Bioinformatics*, vol. 28, no. 12, pp. i137–i146, 2012.

[16] S. Lauritzen, *Graphical Models*, Graphical Models, Oxford University Press, Oxford, UK, 1996.

[17] O. Banerjee, L. El Ghaoui, and A. D'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[18] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[19] S. Kim, S. Oesterreich, S. Kim, Y. Park, and G. C. Tseng, "Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization," *Biostatistics*, vol. 18, no. 1, pp. 165–179, 2017.

[20] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[21] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[22] C. Desmedt, F. Piette, S. Loi et al., "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clinical Cancer Research*, vol. 13, no. 11, pp. 3207–3214, 2007.

[23] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[24] X. Wang, Y. Lin, C. Song, E. Sibille, and G. C. Tseng, "Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder," *BMC Bioinformatics*, vol. 13, no. 1, article 52, 2012.

[25] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[26] M. Farooqui, L. R. Bohrer, N. J. Brady et al., "Epiregulin contributes to breast tumorigenesis through regulating matrix metalloproteinase 1 and promoting cell survival," *Molecular Cancer*, vol. 14, no. 1, article 138, 2015.

[27] H. Hu, N. Takano, L. Xiang, D. M. Gilkes, W. Luo, and G. L. Semenza, "Hypoxia-inducible factors enhance glutamate signaling in cancer cells," *Oncotarget*, vol. 5, no. 19, pp. 8853–8868, 2014.

[28] J. Hou, Z. Wang, H. Xu et al., "Stanniocalicin 2 suppresses breast cancer cell migration and invasion via the PKC/Claudin-1-mediated signaling," *PLoS ONE*, vol. 10, no. 4, Article ID e0122179, 2015.

[29] R. Kimura, A. Kasamatsu, T. Koyama et al., "Glutamate acid decarboxylase 1 promotes metastasis of human oral cancer by β-catenin translocation and MMP7 activation," *BMC Cancer*, vol. 13, article 555, 2013.

[30] M. Martínez-Armenta, S. Díaz de León-Guerrero, A. Catalán et al., "TGFβ2 regulates hypothalamic Trh expression through the TGFβ inducible early gene-1 (TIEG1) during fetal development," *Molecular and Cellular Endocrinology*, vol. 400, pp. 129–139, 2015.

[31] M. Nedergaard, T. Takano, and A. J. Hansen, "Beyond the role of glutamate as a neurotransmitter," *Nature Reviews Neuroscience*, vol. 3, no. 9, pp. 748–755, 2002.