Differential Network Analysis based on Omic Data for Cancer Biomarker Discovery

Yiming Zuo

Dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Electrical Engineering

Guoqiang Yu, Chair

Habtom W. Ressom

Yue Wang

Thomas Charles Clancy

Wenjing Lou

April 28th, 2017

Arlington, Virginia

Keywords: differential expression analysis, differential network analysis, cancer biomarker discovery

Differential Network Analysis based on Omic Data for Cancer Biomarker Discovery

Yiming Zuo

ABSTRACT

Recent advances in high-throughput technique enables the generation of a large amount of omic data such as genomics, transcriptomics, proteomics, metabolomics, glycomics etc. Typically, differential expression analysis (e.g., student's *t*-test, ANOVA) is performed to identify biomolecules (e.g., genes, proteins, metabolites, glycans) with significant changes on individual level between biologically disparate groups (disease cases vs. healthy controls) for cancer biomarker discovery. However, differential expression analysis on independent studies for the same clinical types of patients often led to different sets of significant biomolecules and had only few in common. This may be attributed to the fact that biomolecules are members of strongly intertwined biological pathways and highly interactive with each other. Without considering these interactions, differential expression analysis could lead to biased results.

Network-based methods provide a natural framework to study the interactions between biomolecules. Commonly used data-driven network models include relevance network, Bayesian network and Gaussian graphical models. In addition to data-driven network models, there are many publicly available databases such as STRING, KEGG, Reactome, and ConsensusPathDB, where one can extract various types of interactions to build knowledge-driven networks. While both data- and knowledge-driven networks have their pros and cons, an appropriate approach to incorporate the prior biological knowledge from publicly available databases into data-driven network model is desirable for more robust and biologically relevant network reconstruction.

Recently, there has been a growing interest in differential network analysis, where the connection in the network represents a statistically significant change in the pairwise interaction between two biomolecules in different groups. From the rewiring interactions shown in differential networks, biomolecules that have strongly altered connectivity between distinct biological groups can be identified. These biomolecules might play an important role in the disease under study. In fact, differential expression and differential network analyses investigate omic data from two complementary perspectives: the former focuses on the change in individual biomolecule level between different groups while the latter concentrates on the change in pairwise biomolecules level. Therefore, an approach that can integrate differential expression and differential network analyses is likely to discover more reliable and powerful biomarkers.

To achieve these goals, we start by proposing a novel data-driven network model (i.e., LOPC) to reconstruct sparse biological networks. The sparse networks only contains direct interactions between biomolecules which can help researchers to focus on the more informative connections. Then we propose a novel method (i.e., dwgLASSO) to incorporate prior biological knowledge into data-driven network model to build biologically relevant networks. Differential network analysis is applied based on the networks constructed for biologically disparate groups to identify cancer biomarker candidates. Finally, we propose a novel network-based approach (i.e., INDEED) to integrate differential expression and differential network analyses to identify more reliable and powerful cancer biomarker candidates. INDEED is further expanded as INDEED-M to utilize omic data at different levels of human biological system (e.g., transcriptomics, proteomics, metabolomics), which we believe is promising to increase our understanding of cancer. Matlab and R packages for the proposed methods are developed and available at Github (https://github.com/Hurricaner1989) to share with the research community.

Differential Network Analysis based on Omic Data for Cancer Biomarker Discovery

Yiming Zuo

GENERAL AUDIENCE ABSTRACT

High-throughput technique such as transcriptomics, proteomics and metabolomics is widely used to generate 'big' data for cancer biomarker discovery. Typically, differential expression analysis is performed to identify cancer biomarkers. However, discrepancies from independent studies for the same clinical types of samples using differential expression analysis are observed. This may be attributed to that biomolecules such as genes, proteins and metabolites are members of strongly intertwined biological pathways and highly interactive with each other. Without considering these interactions, differential expression analysis could lead to biased results. In this dissertation, we propose to identify cancer biomarker candidates using network-based approaches. We start by proposing a novel data-driven network model (i.e., LOPC) to reconstruct sparse biological networks. Then we propose a novel method (i.e., wgLASSO) to incorporate prior biological knowledge from public available databases into purely data-driven network model to build biologically relevant networks. In addition, a novel differential network analysis method (i.e., dwgLASSO) is proposed to identify cancer biomarkers. Finally, we propose a novel network-based approach (i.e., INDEED) to integrate differential expression and differential network analyses. INDEED is further expanded as INDEED-M to utilize omic data at different levels of human biological system (e.g., transcriptomics, proteomics, and metabolomics) to identify cancer biomarkers from a systems biology perspective. Matlab and R packages for the proposed methods are developed and shared with the research community.

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisors, Dr. Habtom W. Ressom and Dr. Guoqiang Yu, for their guidance and support during my PhD study. Dr. Ressom has provided me an enjoyable working environment at Georgetown University with talented researchers having diverse backgrounds. I feel relaxed working with him during the past five years. Dr. Yu offers me advice and inspirations to overcome numerous obstacles in both research and life.

Besides my advisors, my sincere gratitude also goes to the rest of my PhD committee, Dr. Yue Wang, Dr. T. Charles Clancy and Dr. Wenjing Lou, for their patience and feedback on my dissertation work. In particular, I want to thank Dr. Wang. I still remember the time when I accepted the GRA offer from him five years ago. We had nice conversations, which encouraged me to start my PhD at Virginia Tech.

I am grateful to Dr. Ruijiang Li at Stanford University. The one-year visiting student researcher time has broadened my horizons from various perspectives and contributed to two publications that was included in this dissertation. I am also grateful to Jelle Ferwerda, CEO & Founder of LogicNets Inc. During the past two years, I have been actively involved in two projects at LogicNets Inc. and learnt a lot by working in his team.

My gratitude is extended to other colleagues and collaborators, Bin Zhou, Dr. Jinlian Wang, Dr. James Li, Dr. Mahlet G Tadesse, Rency S. Varghese, Dr. Cristina Di Poto, Dr. Yi Zhao, Dr. Tsung-Heng Tsai and Minkun Wang from Georgetown University, Dr. Bai Zhang, Dr. Jinghua Gu, Dr. Ye Tian, Dr. Niya Wang, Dr. Xi Chen, Dr. Xiao Wang, and Xu Shi from

Virginia Tech, and Dr. Yi Cui, Dr. Jia Wu and Dr. Hongcheng Liu from Stanford University for their generous help in both research and life.

Finally, I want to express special thanks to my parents, Zhencheng Zuo and Lixin Wang, for their love, encouragement and continuous support throughout my PhD study and my life in general.

Table of contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike information criterion
ANOVA	Analysis of variance
AUC	Area under the curve
BIC	Bayesian information criterion
BN	Bayesian network
CIA	Co-inertia analysis
DE	Differential expression
DEA	Differential expression analysis
DN	Differential network
dwgLASSO	Differentially weighted graphical LASSO
FDR	False Discovery Rate
FN	False negative
FP	False positive
GC-LS	Gas chromatography coupled with mass spectrometry
GGM	Gaussian graphical model
gLASSO	Graphical LASSO
GO	Gene ontology
GSEA	Gene set enrichment analysis
HCC	Hepatocellular carcinoma

INDEED	Integrated differential expression and differential network analysis
INDEED-M	INDEED – multi-omic
KDDN	Knowledge-fused differential dependency network
LASSO	Least absolute shrinkage and selection operator
LC-MS	Liquid chromatography coupled with mass spectrometry
LOPC	Low order partial correlation
MAP	Maximum a posteriori
MCIA	Multiple co-inertia analysis
NS	Neighbor selection
PC	Partial correlation
PPI	Protein-protein interaction
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristic
SAM	Significance analysis of microarrays
sCCA	Sparse canonical correlation analysis
sGCCA	Sparse generalized canonical correlation analysis
SIM	Selected ion monitoring
sPLS	Sparse partial least square
TCGA	The cancer genome atlas
wgLASSO	Weighted graphical LASSO

1 Introduction

1.1 Background and motivation

Recent advances in high-throughput technique enables the generation of a large amount of omic data at different levels of human biological system such as genomics, transcriptomics, proteomics, metabolomics, etc. With these omic data available, understanding the mechanism of cancer requires to identify the difference between disease cases and healthy controls on different biomolecules (i.e., gene, protein, metabolite, etc.). Typically, differential expression analysis (e.g., student's *t*-test, LASSO, etc.) is performed to discover biomolecules with significant changes on individual level between biologically disparate groups. However, independent studies for the same clinical types of patients often lead to different sets of significant biomolecules and had only few in common [1]. This may be attributed to the fact that biomolecules are members of strongly intertwined biological pathways and are highly interactive with each other. Without considering the interactions between them, we may easily get biased result. In addition, to gain deep insights of the roles of the identified significant biomolecules from differential expression analysis, we need to study them in the context of biological pathways they are involved in. Methods such as gene set enrichment analysis (GSEA) test the significance of the overall changes for a list of pre-defined gene sets and pathways [2]. These pre-defined gene sets and pathways are based on the known knowledge, it will be desirable to discover novel knowledge from the omic data by considering the interactions between different biomolecules.

Network-based methods provide a natural framework to study the interactions among biomolecules [3]. A network is constructed to mimic the human biological system as shown in Figure 1.1, in which nodes represent biomolecules and edges indicate the interactions between them. From the reconstructed network, hypotheses can be generated to identify more reliable and biologically relevant biomarker candidate for cancer studies. For example, efforts have been made to integrate protein-protein interaction (PPI) networks with differential expression analysis to identify connected gene biomarker candidates from the PPI network (i.e., sub-networks) for breast cancer study [4]. The authors reported that the identified sub-networks are more biologically relevant to breast cancer and more reproducible when tested on independent datasets. One drawback for this kind of work comes from the assumption that the same PPI network from publicly available databases are used in biologically disparate groups, while in reality rewiring of biological networks usually happens due to the disease state. To solve this problem, group-specific networks should be built and compared to investigate the hidden rewiring patterns between different groups. This is the idea of differential network analysis. Recently, we have observed a shift in interest from differential expression analysis to differential network analysis [5-7]. In a differential network, the connection represents a statistically significant change in the pairwise association between two biomolecules in different groups as shown in Figure 1.2. The conventional way to measure the pairwise association between two biomolecules is based on Pearson's or Spearman's correlation (ρ_{ij}). A connection is built in the differential network when $|\rho_{ij}^{(1)} - \rho_{ij}^{(2)}|$ is statistically significant away from zero, where the superscript indicates the group index. A drawback for using correlation to measure the pairwise association is that correlation confounds direct and indirect associations [8]. For example, a strong correlation between x_1 and x_2 as well as x_2 and x_3 (direct associations) may lead to a relatively weak but still significantly strong correlation between x_1 and x_3 (indirect association) as shown in Figure 2.1. When the number of biomolecules is large, correlation tends to generate over-complicated networks, impacting the selection of reliable biomarker candidates in the subsequent analysis. Considering this, refined measurements that can distinguish direct and indirect associations will be helpful in generating a sparse differential network that benefit both network visualization and reliable biomarker candidate selection.

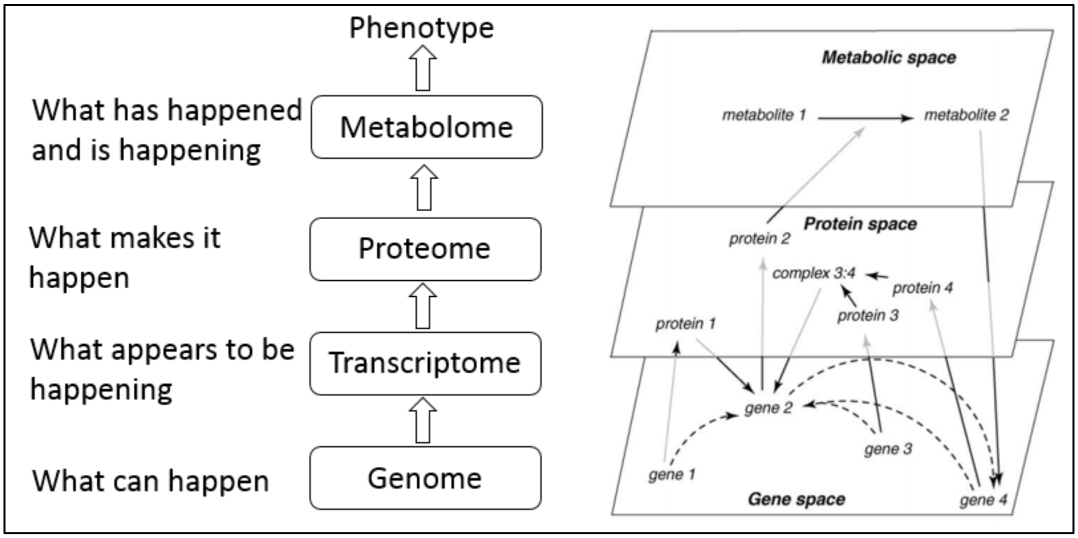


Figure 1.1 The omic cascades for human biological system.

In additional to data-driven network models (e.g., correlation), there are many publicly available databases such as STRING (http://string-db.org/), KEGG (http://www.genome.jp/kegg/), Reactome (http://www.reactome.org/), and ConsensusPathDB (http://consensuspathdb.org/), where one can extract various types of interactions including protein-protein, signaling, and gene regulatory interactions [9-12]. Biological networks reconstructed from these databases have been reported useful. For example, Chuang *et al.* reconstructed protein-protein interaction (PPI) network from multiple databases to help identify markers of metastasis for breast cancer studies using gene expression data [4]. However, databases are far from being complete. Networks constructed purely based on the databases have a large number of false negatives. What's more, databases are seldom specific to a certain disease, so the interactions that exist in the databases may not be reflective of the patient population under study. In contrast, data-driven models are likely to have a large number of false positives due to background noise. Considering this, an appropriate approach to integrate the prior biological knowledge from databases into group-specific data-driven networks is desirable to identify biologically relevant biomarker candidates.

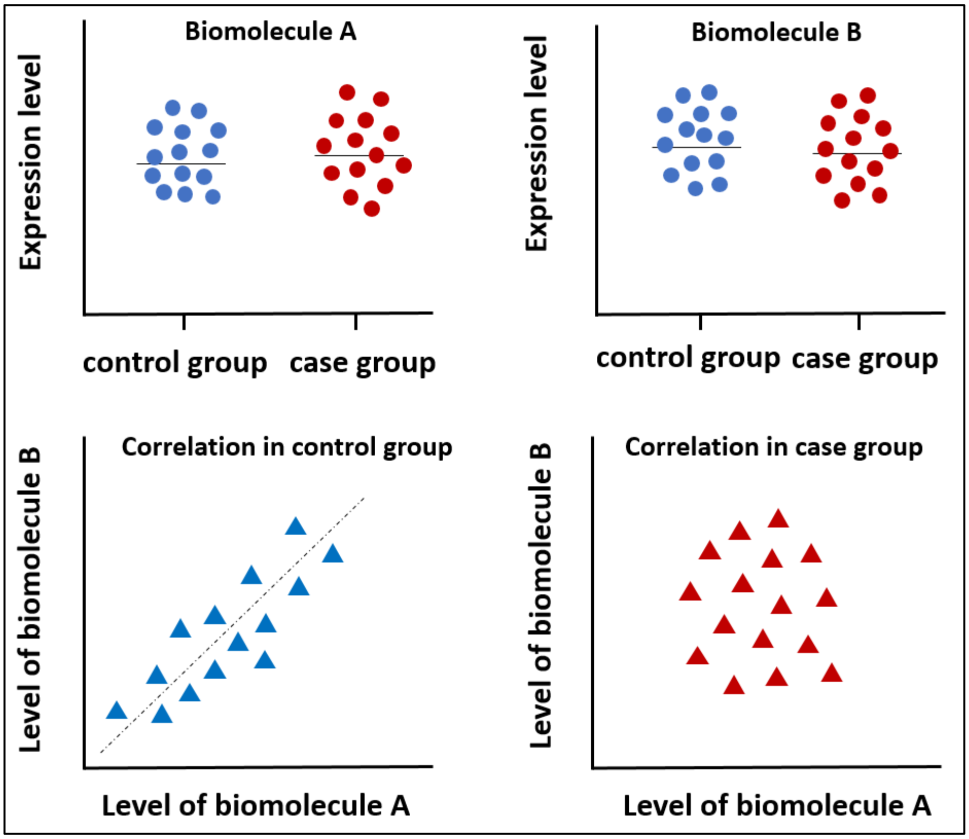


Figure 1.2 Example of a pair of biomolecules with unchanged individual expression levels but changed pairwise associations between case and control groups. Each blue dot corresponds to a control subject while each red dot corresponds to a case subject. Whereas the correlation between biomolecules A and B in the control group is high, this is not the case in the case group.

Given a differential network, an intuitive way to select biomarker candidate is based on node degree (i.e., the number of connections for each node). The assumption is that biomolecules that have a strongly altered connectivity between biologically disparate groups might play an important role in the disease under study [13]. While reasonable, this approach does not consider the change of individual biomolecule between different groups. In fact, differential expression and differential network analyses investigate omic data from two complementary levels: the former focuses on the change for individual biomolecule between different groups while the latter concentrates on the change in biomolecular pairs. Therefore, an approach to integrate differential expression and differential network analyses is likely to lead to more reliable biomarker candidates by considering the difference on a single biomolecule and biomolecular pair levels. Additionally, if multi-omic data for the same set of samples are available (e.g., transcriptomics, proteomics, metabolomics), an approach that can simultaneously integrate differential expression and differential network analyses, and utilize the information provided from biomolecules at different levels of human biological system (e.g., genes, proteins, metabolites) is promising to increase our understanding of cancer and identify more powerful biomarker candidates.

1.2 Objectives and problem statement

As stated in Section 1.1, our goal is to develop a novel method to integrate differential expression and differential network analyses based on single and multi-omic data for cancer biomarker discovery. In order to do that, we started by developing a novel sparse network reconstruction method LOPC to quantify the pairwise association between two biomolecules using partial correlation. Compared with correlation, partial correlation is able to remove the effect of indirect association from other biomolecules given a biomolecular pair. As a result, network built by LOPC is more sparse and the connections are more informative (i.e., direct association). We then developed a novel method wgLASSO to incorporate prior biological knowledge into partial correlation based data-driven networks. The prior biological knowledge was derived from publicly available databases such as STRING (http://string-db.org/), KEGG (http://www.genome.jp/kegg/), etc. The resulting prior biological knowledge incorporated, partial correlation based network is expected to be more beneficial in network visualization and biologically relevant biomarker candidates discovery. Based on wgLASSO, we developed a novel differential network analysis method dwgLASSO, in which differential expression analysis is performed to preselect cancer biomarker candidates, group-specific networks are built using wgLASSO, and the topological changes between the group-specific networks are investigated to identify the final cancer biomarker candidates. By performing differential network analysis with the help of prior biological knowledge, the cancer biomarker candidates selected by dwgLASSO are more robust across independent studies and lead to better performance on predicting the survival time of breast cancer patients based on microarray data when compared with typical differential expression analysis [14].

With all these works and experience in reconstructing sparse biological network, incorporating prior biological knowledge into data-driven network model, and differential network analysis, we went forward to propose a novel method INDEED to integrate differential expression and differential network analyses for cancer biomarker discovery. This is achieved by 1), performing differential expression analysis to obtain *p*-values for each biomolecule based on the change on individual biomolecule between different groups; 2), building a differential network to investigate the change in partial correlation for a given biomolecular pair between different groups; 3), computing a activity score to consider both *p*-value from individual biomolecule level and the connections in the differential network from biomolecular pairs level; 4), prioritizing biomolecules for biomarker candidate selection based on their activity scores. We applied INDEED on real proteomic and glycomic data generated by liquid chromatography coupled with mass spectrometry for hepatocellular carcinoma (HCC) biomarker discovery study. For each omic data, we used one dataset to select biomarker candidates, built a disease classifier and evaluated the performance of the classifier on an independent dataset. The biomarker candidates, selected by INDEED, were more reproducible across independent datasets, and led to a higher classification accuracy in predicting HCC cases and cirrhotic controls compared with those selected by separate differential expression and differential networks analyses. We also extended INDEED to apply for multi-omic data where biomolecules at different levels of human biological system (e.g., genes, proteins, metabolites) on the same set of samples are available. Matlab and R packages for LOPC, wgLASSO, dwgLASSO and INDEED are uploaded on Github account (https://github.com/Hurricaner1989) to share with the research community.

1.3 Organization of the dissertation

The remainder of this research proposal is organized as follows. In Chapter 2, we introduce a novel method LOPC to reconstruct biological network using partial correlation, show its performance compared with other competing network reconstruction methods based on simulation data, and discuss its application on metabolomic liver cancer data. In Chapter 3, we introduce a novel method wgLASSO to incorporate prior biological knowledge into partial correlation based data-driven network model and compare its performance with other competing purely data-driven methods based on simulation data. In Chapter 4, we introduce a novel differential network analysis method dwgLASSO and discuss its application in predicting survival time for breast cancer patients using microarray data, and classification between HCC tumors and non-tumorous liver tissues on TCGA RNA-seq data. In Chapter 5, we introduce a novel method INDEED to integrate differential expression and differential network analyses and present its performance with separated differential expression and differential network analyses in classification between HCC and liver cirrhotic patients based on proteomic and glycomic data, and predicting survival time for breast cancer patients using microarray data. We also extended INDEED for multi-omic data integration for cancer biomarker discovery. In Chapter 6, we make a summary of our contributions in this dissertation, point out the directions for future works, and provide biography and publication records.

2 Biological network reconstruction using low order partial correlation

2.1 Introduction

Systems biology is a rapidly developing field that gives insights that genes and proteins do not work in isolation in complex diseases such as cancer. To better understand the mechanisms of these diseases, different omic studies (e.g., transcriptomics, proteomics, metabolomics) need to be assembled to take advantage of the complementary information and to investigate how they complement each other. One major challenge in the field of systems biology is reconstructing biological networks, such as gene regulatory network, protein-protein interaction network or metabolic network using high-throughput omic data. Generally speaking, network reconstruction methods can be divided into two groups depending on whether the resulting networks are directed graphs or undirected graphs.

Bayesian network (BN) is the most popular method for directed network reconstruction [15]. BN is a probabilistic graphical model where nodes represent biomolecules such as genes, proteins or metabolites and edges denote conditional dependence relationship. It models the biological networks as directed acyclic graphs. However, in reality, cyclic network structures such as feedback loops are ubiquitous in biological systems and are in many cases associated with specific biological properties [16]. Considering this, the assumption of acyclic structure behind BN is limiting.

In contrast, undirected network reconstruction methods circumvent the problems of inferring cyclic network structures, and therefore are more promising in my opinion. Relevance network is one typical undirected network model [17, 18]. It uses correlation or mutual information to measure the “relevance” between biomolecules and sets a hard threshold to connect high relevant pairs. Relevance network has extensive application due to its simplicity and easy implementation. However, its drawback becomes significant when the variable number increases: it confounds direct and indirect associations [8]. While direct association represents the pure association between two biomolecules, indirect association indicates the induced association due to others. For example, Figure 2.1B illustrates that given three biomolecules (x_1, x_2 and x_3), a strong correlation between x_1 and x_2 as well as x_2 and x_3 (direct association) may lead to a relatively weak but still significantly large correlation between x_1 and x_3 (indirect association). As a result, when the number of biomolecules is large, relevance network tends to generate over-complicated networks that contain overwhelming spurious edges (i.e., false positives).

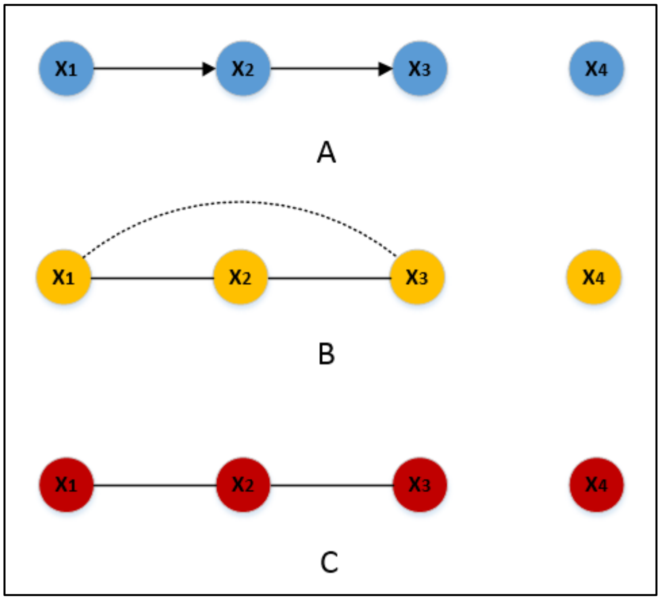


Figure 2.1 Correlation confounds direct and indirect associations while partial correlation does not. (A) The true network from the model. (B) The network inferred based on correlation. The dot line represents the spurious edge due to the indirect association. (C) The network inferred based on partial correlation.

Instead of using correlation or mutual information, partial correlation can distinguish between direct and indirect associations. The partial correlation between x_1 and x_2 given a set of other variables $X^+ = \{x_3, x_4, \dots, x_p\}$ is defined as the correlation between the residuals resulting from the linear regression of x_1 with X^+ and that of x_2 with X^+. To be specific, the partial correlation between x and y given z ($r_{xy \cdot z}$) is obtained by first regressing x on z and y on z separately and then calculating the correlation between the residuals of the models for x and y shown as follows:

$$r_{xy \cdot z} = cor(\varepsilon_x, \varepsilon_y), \tag{2.1}$$

$$\varepsilon_x = x - a - b \times z \tag{2.1-1}$$

$$\varepsilon_y = y - c - d \times z \tag{2.1-2}$$

where ε_x, ε_y are the residuals of x and y after regressing on z; a, b, c, d are regression coefficients.

One widely used method for undirected network inference with partial correlation is Gaussian graphical models (GGMs) [19]. For an undirected graph with p biomolecules, GGM calculates the partial correlation between each pair of nodes conditional on all other $p - 2$ biomolecules. But GGM requires to obtain the inverse covariance matrix. In a 'small n, large p' scenario in omic data, covariance matrix is singular. Considering this, methods based on low order partial correlation have been proposed [20-23]. Low order partial correlation between two biomolecules is obtained only conditional on a subset rather than all other biomolecules. If only zero-th order and first order partial correlations are considered, the resulting undirected graph is called 0-1 graph [24]. 0-1 graph has the advantage that it can be efficiently estimated from small sample-size data, but it fails to infer complex network structure (e.g., cyclic structures) as seen in

Figure 2.2D. The reason is that in Figure 2.2A, x_1 can reach x_4 through either x_2 or x_4, so only conditioning on one of them (0-1 graph only calculates up to the first order partial correlation) is not enough to remove the indirect association between x_1 and x_4. In [20], de la Fuente *et al.* proposed to calculate up to the second order partial correlation to take into account more complex network structure while trying to keep the computational complexity still manageable. However, calculating the second order partial correlation for a typical microarray dataset usually involves several thousands of genes, which is computationally expensive. This limits the application of the method proposed by de la Fuente *et al.* to reconstruct large biological networks. In addition, their method sets threshold empirically without any statistical support.

In view of this, we proposed a more efficient and mathematically sound algorithm (i.e., LOPC) to reconstruct biological networks by calculating partial correlation from zero-th order up to the second order [8]. For a given dataset with *p* biomolecules, we first compute the zero-th order and first order partial correlation for each biomolecular pair. Then, we calculate the second order partial correlation only when both the zero-th order and first order partial correlations are significantly different from zero. With this step, the efficiency of LOPC is largely increased since it excludes most of the possible pairs before calculating the second order partial correlation. Furthermore, we use Fisher's z transformation to create test statistics to set a reasonable threshold. To take into account multiple testing, we control the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure. Simulation results show that LOPC works well under various conditions and the false positives for the inferred network are significantly reduced compared with conventional correlation, 0-1 graph and GGM. We then apply LOPC on a real metabolomics dataset, the result validates the performance of LOPC and shows its potential in discovering novel biomarkers.

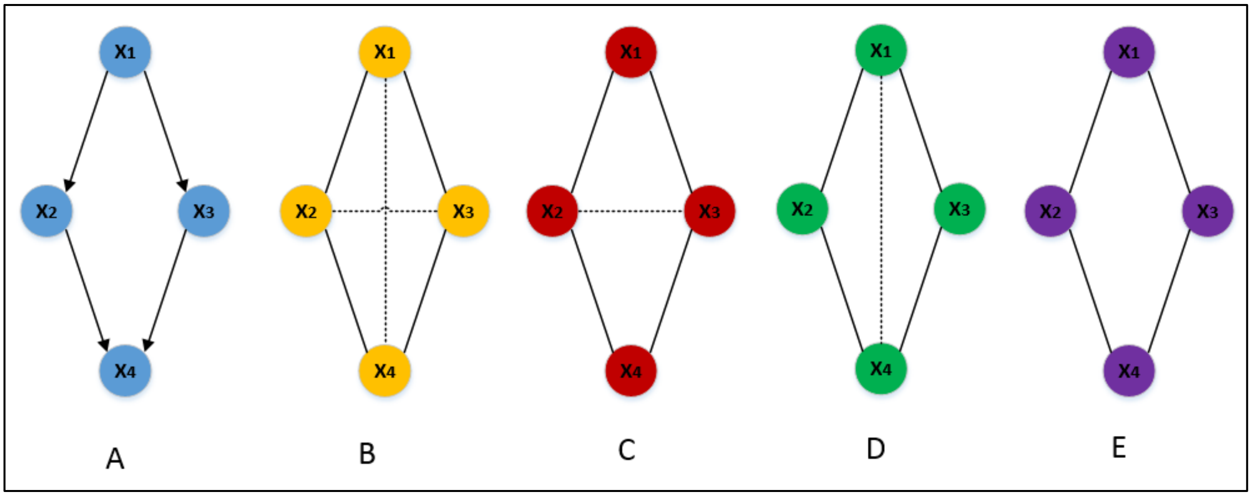


Figure 2.2 Cyclic structure networks inferred based on correlation, GGM, 0-1 graph and low order partial correlation. (A) The true network from the model. (B) Network inferred based on correlation: the dot lines represent the spurious edges. (C) Network inferred based on GGM: by only conditioning on the $(p-2)$-th order (i.e., second order in this model), it is insufficient to uncover the relationships between variables faithfully. (D) Network inferred based on 0-1 graph (up to first order): by only conditioning on up to first order, the indirect association between x_1 and x_4 cannot be removed since there are two paths from x_1 to x_4 either through x_2 or x_3. (E) Network inferred based on low order partial correlation (up to second order): the connections in A are faithfully uncovered.

In the following, we discuss different undirected network reconstruction methods based on correlation, GGM and low order partial correlation in Section 2.2. Also, we introduce test statistics for correlation and partial correlation methods. Then, in Section 2.3, we discuss the proposed algorithm, LPOC. Section 2.4 presents two simulation datasets and a real metabolomics dataset to evaluate the performance of LOPC. Finally, Section 2.5 summarizes our work.

2.2 Low order partial correlation

2.2.1 Undirected network construction methods

Consider p random variables x_1, x_2, ..., x_p, denoted by $X = \{x_1, x_2, \dots, x_p\}$, which may represent gene expression, protein expression or metabolite intensity. Suppose the covariance matrix of X is $\mathbf{\Sigma}$, the correlation between x_i and x_j is defined as:

$$\rho_{ij} = \frac{\mathbf{\Sigma}_{ij}}{\sqrt{\mathbf{\Sigma}_{ii}\mathbf{\Sigma}_{jj}}} \tag{2.2}$$

In a correlation-based network, x_i and x_j are considered to be connected when $\rho_{ij} \neq 0$ (i.e., its estimate r_{ij} is significantly different from zero).

One common criticism for the above correlation-based network is that it yields too many spurious edges since correlation confounds direct and indirect associations. Let's consider an example where $X = \{x_1, x_2, x_3, x_4\}$ and the relationships between x_1, x_2, x_3 and x_4 are modeled as $x_1 = s + \epsilon_1$, $x_2 = \lambda \times x_1 + \epsilon_2$, $x_3 = \mu \times x_2 + \epsilon_3$, $x_4 = \epsilon_4$ assuming $s \sim N(0, \sigma_s^2)$, denoting the signal; ϵ_1, ϵ_2, ϵ_3, $\epsilon_4 \sim N(0, \sigma_n^2)$, denoting the independent and identically distributed (i.i.d.) noise; signal and noise are independent; λ, μ are non-zero constants. Figure 2.1A represents the above relationships and the arrow directions are assigned manually to represent causality. In this model, the relationships between x_1 and x_2, x_2 and x_3 are direct associations while x_1 and x_3 are indirectly related. Figure 2.1B shows that the undirected network inferred based on correlation confounds direct and indirect associations, thus leading to a spurious edge or false positive (i.e., the edge between x_1 and x_3).

In contrast, GGM removes the linear effect of all remaining $p - 2$ variables when calculating the partial correlation between two variables conditional on all other variables.

Suppose X follows a multivariate Gaussian distribution, R and Q are two subsets of X where $R = \{x_i, x_j\}$ and $Q = X \setminus R$. The conditional covariance matrix of R given Q can be computed as follows when $\mathbf{\Sigma}_{QQ}$ is nonsingular:

$$\mathbf{\Sigma}_{R|Q} = \mathbf{\Sigma}_{RR} - \mathbf{\Sigma}_{RQ}\mathbf{\Sigma}_{RR}^{-1}\mathbf{\Sigma}_{QR} \tag{2.3}$$

where the covariance matrix of X is $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{RR} & \mathbf{\Sigma}_{RQ} \\ \mathbf{\Sigma}_{QR} & \mathbf{\Sigma}_{QQ} \end{bmatrix}$.

Similarly, the precision matrix of X (the inverse of $\mathbf{\Sigma}$) can be represented as:

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{\Theta}_{RR} & \mathbf{\Theta}_{RQ} \\ \mathbf{\Theta}_{QR} & \mathbf{\Theta}_{QQ} \end{bmatrix} \tag{2.4}$$

where $\mathbf{\Theta}_{RR} = \left(\mathbf{\Sigma}_{RR} - \mathbf{\Sigma}_{RQ}\mathbf{\Sigma}_{QQ}^{-1}\mathbf{\Sigma}_{QR}\right)^{-1}$ [25].

Suppose $\mathbf{\Theta}_{RR} = \begin{bmatrix} \theta_{ii} & \theta_{ij} \\ \theta_{ji} & \theta_{jj} \end{bmatrix}$, then from Equation 2.3, the conditional covariance matrix $\mathbf{\Sigma}_{R|Q}$ can be obtained as:

$$\mathbf{\Sigma}_{R|Q} = \mathbf{\Theta}_{RR}^{-1} = \frac{1}{\det(\mathbf{\Theta}_{RR})}\begin{bmatrix} \theta_{jj} & -\theta_{ij} \\ -\theta_{ji} & \theta_{ii} \end{bmatrix} \tag{2.5}$$

Under the Gaussian distribution, partial correlation and conditional correlation are equivalent. A proof involving three variables is shown in the Appendix A. For a more general proof, one can refer to [26, 27]. Once the precision matrix is known, the partial correlation coefficient between x_i and x_j conditional on all other variables can be computed as:

$$\rho_{ij\cdot Q} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \tag{2.6}$$

In a GGM-based network, x_i and x_j are considered to be connected when $\rho_{ij\cdot Q} \neq 0$ (i.e., its estimate $r_{ij\cdot Q}$ is significantly different from zero).

By removing the effect of all other variables, GGM can distinguish between direct and indirect associations as seen in Figure 2.1C. However, it requires that the covariance matrix be full rank for a well-defined matrix inversion, so the sample size should be at least as large as the number of variables. This poses a challenge for most omic datasets, which typically involve thousands of variables but much less number of samples.

Rather than conditioning on all other variables, low order partial correlation conditions on only a few of them. The order of the partial correlation coefficient is determined by the number of variables it conditions on. The advantage of using low order partial correlation relies on a recursive equation (i.e., a higher order partial correlation can be computed from its preceding order) [28].

For $X = \{x_1, x_2, x_3, x_4\}$ modeled in Figure 2.2, without loss of generality, we assume $\sigma_s^2 = \sigma_n^2 = 1$, then the covariance matrix $\mathbf{\Sigma}$ is:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \lambda & \lambda\mu & 0 \\ \lambda & \lambda^2 + 1 & (\lambda^2 + 1)\mu & 0 \\ \lambda\mu & (\lambda^2 + 1)\mu & (\lambda^2 + 1)\mu^2 + 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.7}$$

From Equation 2.3, the conditional covariance matrix of $\{x_1, x_2\}$ given x_3 is:

$$\mathbf{\Sigma}_{1,2|3} = \frac{1}{(\lambda^2+1)\mu^2+1} \begin{bmatrix} \mu^2 + 1 & \lambda \\ \lambda & \lambda^2 + 1 \end{bmatrix} \tag{2.8}$$

Since partial correlation is equivalent to conditional correlation under Gaussian distribution, the first order partial correlation between x_1 and x_2 conditional on x_3 can be computed from Equation 2.8:

$$\rho_{12\cdot3} = \frac{\lambda}{\sqrt{(\lambda^2+1)(\mu^2+1)}} \tag{2.9}$$

When the zero-th order partial correlation ρ_{12}, ρ_{13}, ρ_{23} are computed from Equation 2.7 and compared with Equation 2.9, the following relationship exists between zero-th order and the first order partial correlations:

$$\rho_{12\cdot3} = \frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}} \tag{2.10}$$

Equation 2.10 can be generalized so that higher order partial correlation can be calculated from its preceding order. For example, similar equation exists between the first order and the second order partial correlations:

$$\rho_{12\cdot34} = \frac{\rho_{12\cdot3}-\rho_{14\cdot3}\rho_{24\cdot3}}{\sqrt{(1-\rho_{14\cdot3}^2)(1-\rho_{24\cdot3}^2)}} \tag{2.11}$$

Theoretically, in order to obtain the exact undirected graph for p variables, one needs to potentially calculate the partial correlations from zero-th order up to the $(p-2)$-th order [29]. Correlation considers only the zero-th order; GGM consider only the $(p-2)$-th order. It was previously reported that neither of them is sufficient to uncover the conditionally independent relationships between variables [20]. Though the idea behind low order partial correlation is simple, it can serve as a good approximation to the true network as seen in Figure 2.2. In addition, low order partial correlation has the advantage of working well when the sample size is small and the number of variable is large.

If only zero-th order and first order partial correlations are considered, the resulting network is called a 0-1 graph. The network is constructed based on the following rule: the edge between nodes x_i and x_j is connected when all r_{ij} and $r_{ij\cdot k}$ are significantly away from zero, where k considers each possible x_k in $X \setminus \{x_i, x_j\}$.

Similarly, if we calculate up to the second order partial correlation, the resulting network is constructed based on the rule that the edge between nodes x_i and x_j is connected when all r_{ij}, $r_{ij\cdot k}$ and $r_{ij\cdot kq}$ are significantly different from zero, where k and q correspond to every possible x_k and x_q in $X \setminus \{x_i, x_j\}$.

2.2.2 Test statistics

The test statistics for non-zero correlation with a sample size of n is:

$$t(r_{ij}) = \frac{r_{\mathrm{ij}}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}} \tag{2.12}$$

For a zero correlation, $t(r_{ij})$ follows a t-distribution with $n-2$ degrees of freedom. In contrast, the test statistics for non-zero partial correlation can be calculated using the Fisher's z-transformation [30]:

$$z(r_{ij\cdot \mathrm{Q}}) = \frac{1}{2}\ln\left(\frac{1+r_{ij\cdot Q}}{1-r_{ij\cdot Q}}\right) \tag{2.13}$$

where Q corresponds to $X \setminus \{x_i, x_j\}$.

For a zero partial correlation with sample size n, $z(r_{ij\cdot Q})$ is approximately Gaussian distributed with zero mean and $\frac{1}{n-|Q|-3}$ variance, where $|Q|$ represents the number of elements in Q. Given a partial correlation, the two-sided p-value is:

$$p\left(r_{ij\cdot Q}\right) = 2 \times \left\{1 - \phi\left[\sqrt{n - |Q| - 3} \cdot z\left(r_{ij\cdot Q}\right)\right]\right\} \tag{2.14}$$

2.2.3 LOPC algorithm

The proposed LOPC algorithm contains four steps:

1) Calculate the zero-th, first and second order partial correlations;

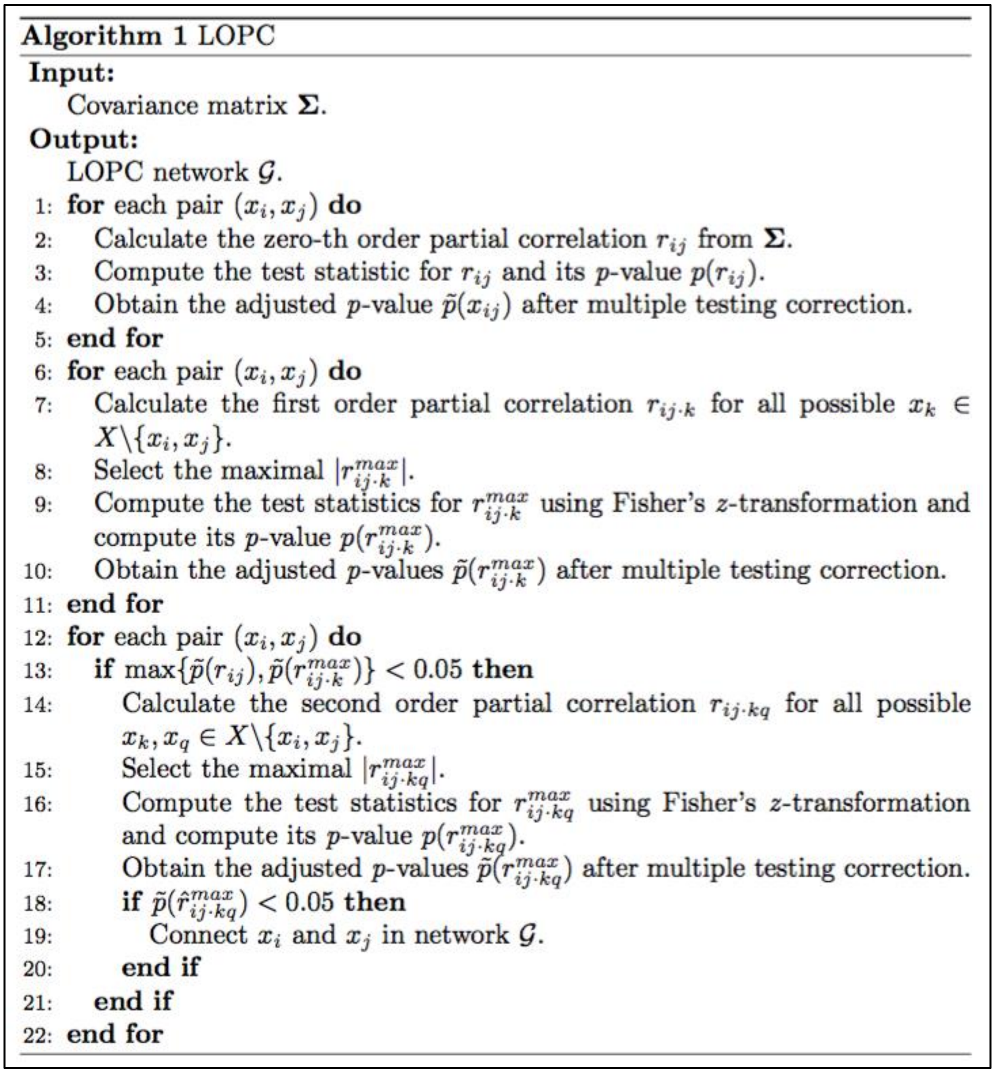
2) Calculate test statistics and corresponding p-values to evaluate the null hypothesis that the corresponding partial correlation is zero;

3) Calculate the adjusted p-values for multiple testing correction;

4) Construct the network.

Among the four steps, most of the computation time is spent on calculating the second order partial correlation $r_{ij\cdot kq}$ since one needs to consider all possible x_k, x_q in $X \setminus \{x_i, x_j\}$. It was previously suggested that the distribution of connections in metabolic, regulatory and protein-protein interaction networks tends to follow a power law [31, 32]. Thus, the resulting networks are very sparse.

Here, we present an efficient algorithm taking advantage of this sparsity property of biological networks. Instead of calculating the second order partial correlation $r_{ij\cdot kq}$ for all possible x_i and x_j, we only calculate those whose corresponding zero-th and first order partial correlations are significantly different from zero. Since the true biological networks are sparse, this step can exclude most of the possible spurious edges before calculating the second order partial correlation. As a result, our proposed LOPC algorithm can dramatically reduce the computational burden. The detailed algorithm is shown below:



Algorithm 1 LOPC

```
Input:
    Covariance matrix $\mathbf{\Sigma}$.
Output:
    LOPC network $\mathcal{G}$.
for each pair $(x_i, x_j)$ do
    Calculate the zero-th order partial correlation $r_{ij}$ from $\mathbf{\Sigma}$.
    Compute the test statistic for $r_{ij}$ and its $p$-value $p(r_{ij})$.
    Obtain the adjusted $p$-value $\tilde{p}(x_{ij})$ after multiple testing correction.
end for
for each pair $(x_i, x_j)$ do
    Calculate the first order partial correlation $r_{ij \cdot k}$ for all possible $x_k \in X \backslash \{x_i, x_j\}$.
    Select the maximal $|r^{max}_{ij \cdot k}|$.
    Compute the test statistics for $r^{max}_{ij \cdot k}$ using Fisher's $z$-transformation and compute its $p$-value $p(r^{max}_{ij \cdot k})$.
    Obtain the adjusted $p$-values $\tilde{p}(r^{max}_{ij \cdot k})$ after multiple testing correction.
end for
for each pair $(x_i, x_j)$ do
    if $\max\{\tilde{p}(r_{ij}), \tilde{p}(r^{max}_{ij \cdot k})\} < 0.05$ then
        Calculate the second order partial correlation $r_{ij \cdot kq}$ for all possible $x_k, x_q \in X \backslash \{x_i, x_j\}$.
        Select the maximal $|r^{max}_{ij \cdot kq}|$.
        Compute the test statistics for $r^{max}_{ij \cdot kq}$ using Fisher's $z$-transformation and compute its $p$-value $p(r^{max}_{ij \cdot kq})$.
        Obtain the adjusted $p$-values $\tilde{p}(r^{max}_{ij \cdot kq})$ after multiple testing correction.
        if $\tilde{p}(\hat{r}^{max}_{ij \cdot kq}) < 0.05$ then
            Connect $x_i$ and $x_j$ in network $\mathcal{G}$.
        end if
    end if
end for
```

2.3 Evaluation of LOPC on simulated and real data

This section presents two numerical simulations (A and B) to reconstruct undirected network based on correlation, GGM, 0-1 graph, and LOPC, as well as one real application of LOPC on a metabolomic dataset.

2.3.1 LOPC on simulated data

In simulation A, we consider an example where variables $X = \{x_1, x_2, x_3, x_4\}$ form a cyclic structure as shown in Figure 2.2A. The relationships between x_1, x_2, x_3, x_4 were modeled as: $x_1 = s + \epsilon_1$, $x_2 = \lambda \times x_1 + \epsilon_2$, $x_3 = \mu \times x_1 + \epsilon_3$, $x_4 = \alpha \times x_2 + \beta \times x_3 + \epsilon_4$ assuming $s \sim N(0, \sigma_s^2)$, denoting the signal; $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \sim N(0, \sigma_n^2)$, denoting the i.i.d. noise; signal and noise are independent; $\lambda, \mu, \alpha, \beta$ are non-zero constants. Without loss of generality, we set $\sigma_s^2 = 1$, $\sigma_n^2 = 0.01$ and $\lambda = \mu = \alpha = \beta = 1$. The resulting covariance matrix for X is:

$$\mathbf{\Sigma} = \begin{bmatrix} 1.01 & 1.01 & 1.01 & 2.02 \\ 1.01 & 1.02 & 1.01 & 2.03 \\ 1.01 & 1.01 & 1.02 & 2.03 \\ 2.02 & 2.03 & 2.03 & 4.07 \end{bmatrix} \quad (2.15)$$

We generated dataset from $N(\mathbf{0}, \mathbf{\Sigma})$ with a sample size $n = 50$ and reconstructed networks based on correlation, GGM, 0-1 graph and LOPC as seen in Figure 2.2.

The correlation, partial correlation for each pair of variables and the corresponding adjusted p-values are shown in Table 2.1.

Table 2.1 Correlation, partial correlation and p-values for each edge.

Edge	**Correlation**		**GGM**		**0-1 graph**		**LOPC**	
	R	**P**	**R**	**P**	**R**	**P**	**R**	**P**
x1x2	0.994	9.00E-47	0.533	9.00E-04	0.278	5.00E-03	0.533	9.00E-04
x1x3	0.996	9.00E-51	0.658	4.00E-06	0.518	2.00E-07	0.658	4.00E-06
x1x4	0.995	1.00E-50	-0.257	0.774	0.33	8.00E-03	-0.257	0.774
x2x3	0.991	5.00E-45	-0.543	7.00E-04	0.148	8.50E-01	0	1
x2x4	0.997	4.00E-53	0.798	1.00E-10	0.704	0	0.798	1.00E-10
x3x4	0.997	7.00E-53	0.754	6.00E-09	0.634	2.00E-12	0.754	6.00E-09

In Figures 2.2B and 2.2C, we see that both correlation and GGM-based networks yield spurious edges. This is because correlation confounds direct and indirect associations, while

GGM are insufficient to uncover the network structure faithfully in this model by only considering the $(p-2)$-th order partial correlation. In fact, from the perspective of probabilistic graphical models, x_1 is a common ancestor of x_2 and x_3, while x_4 is a causal descendent of x_2 and x_3 [33]. Since conditioning on any common causal descendent would introduce a correlation between two variables, there is a dependence estimated between x_2 and x_3 by conditioning on both x_1 and x_4 using GGM.

The resulting networks based on 0-1 graph and LOPC are shown in Figures 2.2D and 2.2E, respectively. For 0-1 graph, since there are multiple paths from x_1 to x_4 either through x_2 or x_3. By only calculating up to the first order partial correlation, it is insufficient to remove the indirect association between x_1 and x_4. However, when we calculate up to the second order partial correlation, the cyclic structure can be faithfully recovered. In fact, Figure 2.2E can be viewed as the result of merging Figures. 2.2B, 2.2C and 2.2D and only keeping common edges.

In simulation B, we consider a more complex structure where ten variables $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ were involved and their relationships were modeled as: $x_1 = s_1 + \epsilon_1$, $x_{10} = s_2 + \epsilon_{10}$, $x_2 = \lambda_1 \times x_1 + \epsilon_2$, $x_3 = \alpha_1 \times x_2 + \epsilon_3$, $x_4 = \alpha_2 \times x_3 + \epsilon_4$, $x_5 = \alpha_3 \times x_4 + \lambda_2 \times x_1 + \epsilon_5$, $x_6 = \alpha_4 \times x_5 + \mu_1 \times x_{10} + \epsilon_6$, $x_7 = \alpha_5 \times x_6 + \epsilon_7$, $x_8 = \alpha_6 \times x_7 + \epsilon_8$, $x_9 = \alpha_7 \times x_8 + \mu_2 \times x_{10} + \epsilon_9$ with $s1, s2 \sim N(0, \sigma_s^2)$, denoting the signal; $\epsilon_1, \epsilon_2, \dots, \epsilon_{10} \sim N(0, \sigma_n^2)$, denoting the i.i.d. noise; signal and noise are independent; λ_1, λ_2, μ_1, μ_2, α_1, α_2, ..., α_7 are non-zero constants. The network structure is shown in Figure 2.3A. In this network, x_1 and x_{10} can be interpreted as regulators while x_2 to x_9 represent genes, proteins or metabolites being regulated. Correspondingly, λ_1, λ_2, μ_1, μ_2 denote the strength of the regulation. Without loss of generality, we set $\sigma_s^2 = 1$, $\sigma_n^2 = 0.01$, all the coefficients (i.e., λ_1, λ_2, μ_1, μ_2, α_1, α_2, ..., α_7) to be 1. The resulting covariance matrix for X is seen in Equation 2.16.

We generated dataset from $N(\mathbf{0}, \mathbf{\Sigma})$ with a sample size $n = 50$ and reconstructed networks based on correlation, GGM, 0-1 graph and LOPC. For the correlation-based network, the number of spurious edges (false positives) increases dramatically with nearly every possible variable pair being connected. In Figures 2.3B to 2.3D, we show the inferred networks based on GGM, 0-1 graph and LOPC.

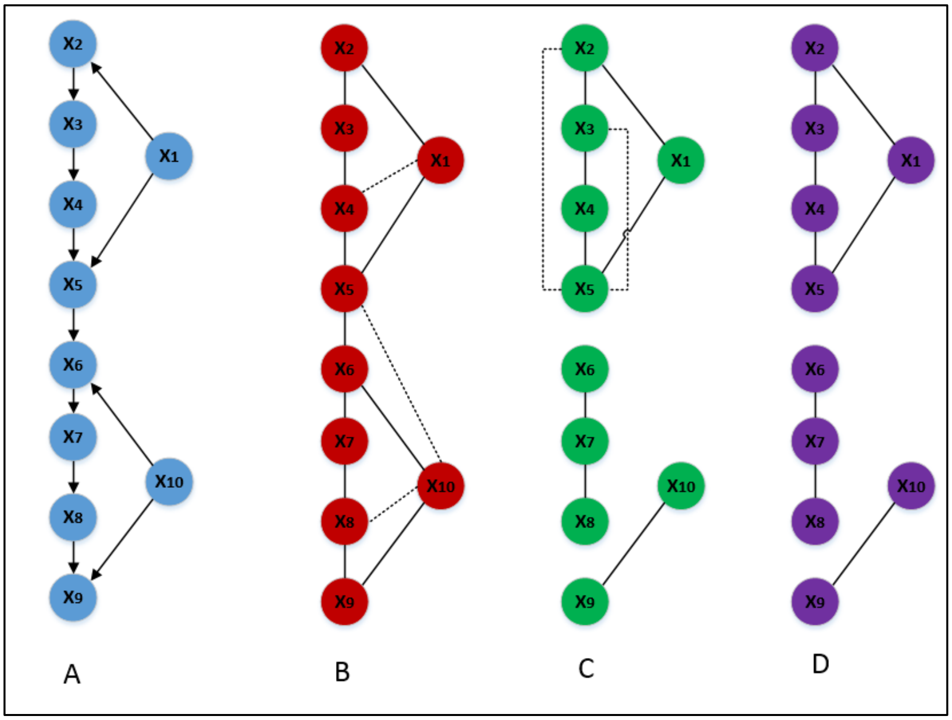


Figure 2.3 Complex structure networks inferred based on GGM, 0-1 graph and LOPC. (A) The true network from the model. (B) Network inferred based on GGM: the dot lines represent the spurious edges. (C) Network inferred based on 0-1 graph (up to first order): by only conditioning on up to first order, the resulting inferred network has similar number of spurious edges (false positives) as that from GGM but has several missed edges (false negatives). (D) Network inferred based on LOPC (up to second order): while the missed edges are inherited from 0-1 graph, calculating up to the second order successfully removes spurious edges in the inferred network.

As shown in Figure 2.3B, GGM-based network yields a few false positives. This is because GGM only considers the $(p-2)$-th order partial correlation. As seen in Figure 2.3C, the 0-1 graph yields similar number of false positives compared with GGM-based network but starts to have missing edges (false negatives). The network inferred with LOPC (Figure 2.3D) also has false negatives, but calculating up to the second order removes all the false positives.

$$\mathbf{\Sigma} = \begin{bmatrix} 1.01 & 1.01 & 1.01 & 1.01 & 2.02 & 2.02 & 2.02 & 2.02 & 2.02 & 0 \\ 1.01 & 1.02 & 1.02 & 1.02 & 2.03 & 2.03 & 2.03 & 2.03 & 2.03 & 0 \\ 1.01 & 1.02 & 1.03 & 1.03 & 2.04 & 2.04 & 2.04 & 2.04 & 2.04 & 0 \\ 1.01 & 1.02 & 1.03 & 1.04 & 2.05 & 2.05 & 2.05 & 2.05 & 2.05 & 0 \\ 2.02 & 2.03 & 2.04 & 2.05 & 4.08 & 4.08 & 4.08 & 4.08 & 4.08 & 0 \\ 2.02 & 2.03 & 2.04 & 2.05 & 4.08 & 5.1 & 5.1 & 5.1 & 6.11 & 1.01 \\ 2.02 & 2.03 & 2.04 & 2.05 & 4.08 & 5.1 & 5.11 & 5.11 & 6.12 & 1.01 \\ 2.02 & 2.03 & 2.04 & 2.05 & 4.08 & 5.1 & 5.11 & 5.12 & 6.13 & 1.01 \\ 2.02 & 2.03 & 2.04 & 2.05 & 4.08 & 6.11 & 6.12 & 6.13 & 8.16 & 2.02 \\ 0 & 0 & 0 & 0 & 0 & 1.01 & 1.01 & 1.01 & 2.02 & 1.01 \end{bmatrix} \tag{2.16}$$

Using a similar model, we generated 100 simulation datasets for varying number of variables and sample sizes and calculated the mean of false positives and false negatives for each method as shown in Table 2.2.

Generally speaking, we expect LOPC to lead to far less number of false positives compared to GGM and 0-1 graph with a possible drawback of selecting a few more false negatives. In real application, this is desirable since one would usually prefer to be confident about the existence of edges already detected, though some edges might be missed. As shown in Table 2.2, when the sample size is slightly larger than the number of variables, LOPC works well, whereas GGM’s performance begins to decline due to the difficulty of inverting singular matrix. To address this, further technique such as graphical lasso has been incorporated into GGM [34].

Table 2.2 Mean of false positives and false negatives for varying number of variables and sample sizes.

Variable Number	Sample Size	Correlation		GGM		0-1 graph		LOPC	
		FP	FN	FP	FN	FP	FN	FP	FN
10	50	28.78	0.11	2.93	0.23	1.95	3.08	0	3.31
20	50	54.91	1.42	1.29	15.06	1.4	10.84	0.09	15.98
	100	57.19	0.41	5.28	2.17	3.64	6.37	0	7.47
50	51	142.55	2.09	7.33	54.61	6.29	20.9	0.42	32.57
	100	145.85	0.33	13.73	9.98	9.73	15.21	0.21	16.86
100	101	299.51	0.47	0	109.98	19.62	30.43	1.56	33.34
	200	299.44	0	39.16	0.78	20.04	29.48	0.83	29.87
500	501	1491.71	0	0	550	89.78	151.32	7.31	171.38
	1000	1495.25	0	150.78	10.34	114.25	99.75	1.35	105.19
1000	1001	2990.13	0	0	1100	181.27	271.78	13.77	301.13

2.3.2 LOPC on real data

In this section, we applied LOPC on a real untargeted metabolomics dataset previously collected and analyzed by our group for hepatocellular carcinoma (HCC) biomarker discovery study [35]. The data were acquired by analysis of sera from 40 HCC cases and 50 patients with liver cirrhosis using liquid chromatography coupled with mass spectrometry (LC-MS). Following preprocessing, a data matrix was obtained with 984 input variables - larger than the sample size 90. We identified 32 metabolites with intensities significantly different between the HCC cases and cirrhotic controls.

Rather than looking into each statistically significant metabolite, we generated undirected network using LOPC after normalization of the preprocessed data matrix. The aim of the normalization is to bring the intensities of the metabolites in both cases and controls to a comparable level. The resulting networks are depicted in Figure 2.4A. We then mapped the 32 statistically significant metabolites onto Figure 2.4A and extracted functional modules which contained multiple metabolites. Two interesting functional modules are shown in Figures 2.4B and 2.4C, respectively, with blue nodes representing the metabolites and white nodes representing non-significant ones. Due to the limitation in metabolite identification, some of the nodes have been assigned multiple putative IDs (e.g., Glycine; Haloperidol decanoate) or have no IDs (unknowns). We see that metabolites connecting with each other tend to be involved in the same chemical reaction and have similar functionalities. The extracted functional modules may help identify other non-significant metabolites that might be missing from the statistical analysis due to subtle differences in ion intensities.

Finally, we evaluated the efficiency of LOPC by randomly sampling various numbers of metabolites from the above dataset to generate 100 undirected networks. We compared the averaged run-time between LOPC and the traditional method to calculate up to the second order partial correlation. While the traditional method calculates the 0th, 1st, and 2nd order partial correlations, LOPC evaluates the outcome of the 1st order partial correlation to determine whether or not the calculation of the 2nd order partial correlation is needed. As shown in Figure 2.5, when the input variable number increases beyond 50, LOPC starts to become more efficient than traditional method. With an input variable number of 200, LOPC can be as 4 times fast as the traditional method. The run-time comparison was performed using a personal computer with an Intel(R) Core(TM) i7-2600 CPU @ 3.4GHz and 16.0 GB RAM.

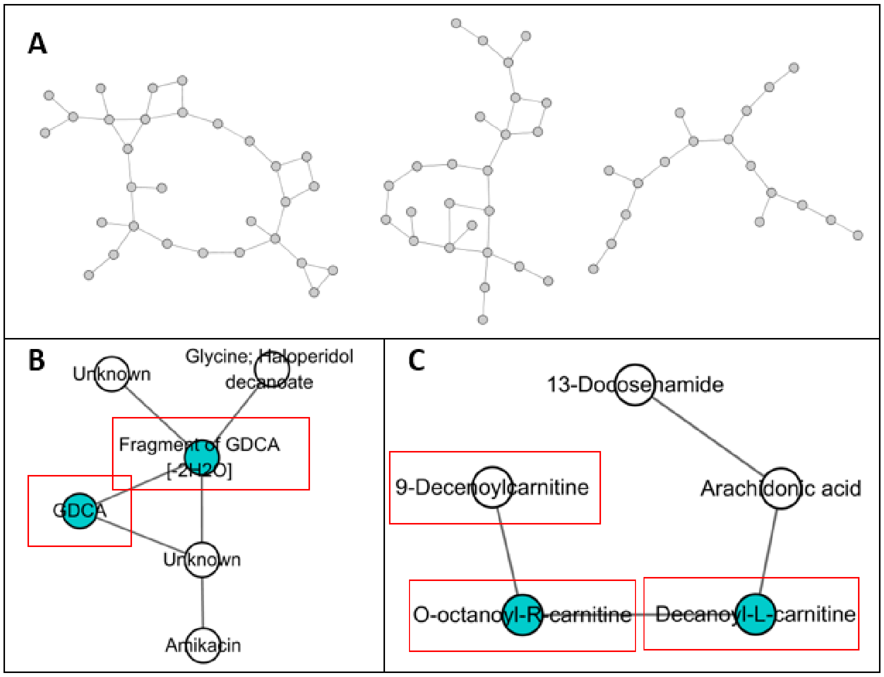


Figure 2.4 Undirected network and functional modules inferred from real data by LOPC. (A) Undirected network encoding the direct associations between different nodes (B-C) Functional modules extracted from the undirected network. Blue nodes represent the candidate biomarkers previously reported. White nodes represent the non-significant ones.

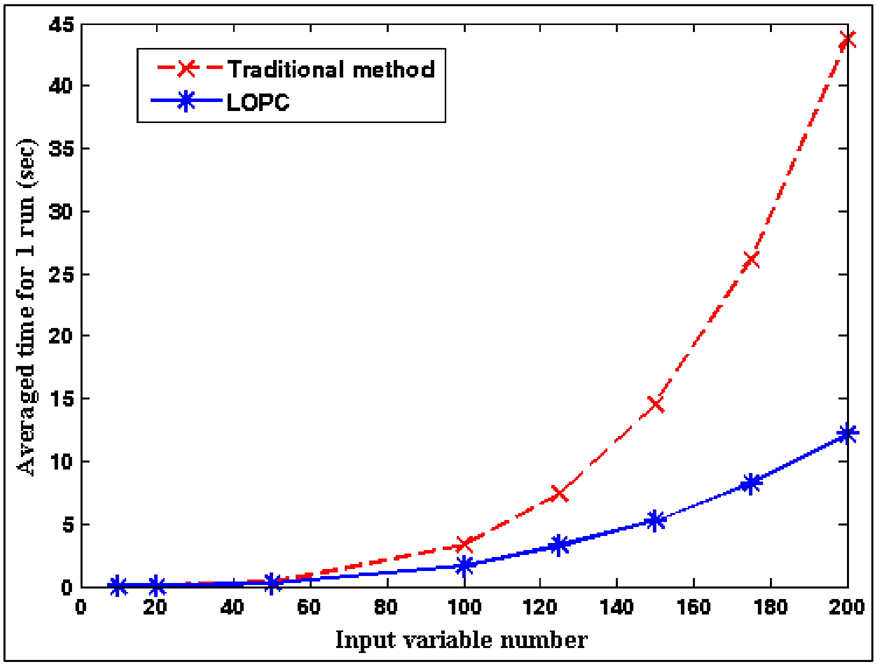


Figure 2.5 Run-time comparison between LOPC and the traditional method in calculating up to the second order partial correlation.

2.4 Conclusion

In Chapter 2, we propose an efficient algorithm LOPC to reconstruct biological networks by calculating up to the second order partial correlation. Compared with other competing undirected network inference methods (correlation, GGM, and 0-1 graph), LOPC offers better solution for inferring networks with less spurious edges (false positives). It also has the advantage of handling well cases that involve a large number of variables but a small sample size. These properties make LOPC a promising alternative to infer from omic datasets relevant gene co-expression, protein-protein interaction and metabolic networks, which may give insights into the mechanisms of complex diseases. A real application on metabolomics dataset validates the performance of LOPC and shows its potential in discovering novel biomarkers. An open source Matlab package is available at Github to share LOPC with the scientific community (https://github.com/Hurricaner1989/LOPC-Matlab-package).

3 Incorporating prior biological knowledge into biological network reconstruction

3.1 Introduction

Recently, Gaussian graphical models (GGMs) have been increasingly applied on biological network inference [36-38]. Similar to Bayesian network, GGMs can remove the effect of indirect associations through estimation of the conditional dependence relationship. At the same time, they are undirected graphs and have no limitation on modeling only acyclic structures. In GGMs, a connection between two nodes corresponds to a non-zero entry in the inverse covariance matrix (i.e., precision matrix), which indicates a conditional dependency between these two nodes given the others. The concept of GGMs dates back to early 1970s when Dempster introduced "covariance selection" problem [39]. The conventional approach to solve GGM problem relies on statistical test (e.g., deviation tests) and forward/backward selection procedure [33]. This is not feasible for high-throughput omic data when the number of genes is ranging from several hundred to thousands while the number of samples are only tens to hundreds since forward/backward selection procedure is computationally too expensive when the variable number is large. In addition, the "small n, large p" scenario for omic data (i.e., sample size is far less than the square of the variable number), makes maximum likelihood estimation (MLE) of precision matrix not to exist because the sample covariance matrix is rank deficient. To deal with these issues, Schäfer *et al.* proposed to combine Moore–Penrose pseudoinverse and bootstrapping technique to approximate the precision matrix [40]. Others applied ℓ_1 regularization to get a sparse network [34, 41, 42]. Taking into account the sparsity property of biological networks and the computational burden of bootstrapping, ℓ_1 regularization methods are preferred. Among them, Meinshausen *et al.* performed ℓ_1 regularized linear regression (i.e., LASSO) for each node to select its "neighbors" [41]. Given all its neighbors, one node is conditional independent with the remaining ones. This 'neighbor selection' approach may face the consistency problem that while gene X is selected as Y's neighbor, gene Y may not be selected as X's neighbor when performing LASSO. We need to decide whether to remove this kind of inconsistent connections based on our experience. A more statistically sound way is to use graphical LASSO, which directly estimates precision matrix by applying ℓ_1 regulation on the elements of the precision matrix to obtain a sparse estimated precision matrix [34, 42]. We are going to pursuit the extension of graphical LASSO in this chapter.

In additional to data-driven network models, there are many publicly available databases such as STRING (http://string-db.org/), KEGG (http://www.genome.jp/kegg/), Reactome (http://www.reactome.org/), and ConsensusPathDB (http://consensuspathdb.org/), where one can extract various types of interactions including protein-protein, signaling, and gene regulatory interactions [9-12]. Biological networks reconstructed from these databases have been reported useful. For example, Chuang *et al.* reconstructed protein-protein interaction (PPI) network from multiple databases to help identify markers of metastasis for breast cancer studies using gene expression data [4]. They overlaid the gene expression value on its corresponding protein in the network and searched for sub-networks whose activities across all patients were highly discriminative of metastasis. By doing this, they found several hub genes related to known breast cancer mutations, while these genes were not found significant by differential gene expression analysis. They also reported that the identified sub-networks are more reproducible between different breast cancer cohorts than individual gene markers. However, databases are far from being complete. Networks constructed purely based on the databases have a large number of false negatives. In addition, databases are seldom specific to a certain disease, so the interactions that exist in the databases may not be reflective of the patient population under study. In contrast, data-driven models are likely to have a large number of false positives due to background noise. Considering this, an appropriate way to integrate the prior biological knowledge from databases and data-driven network model is desirable for more robust and biologically relevant network reconstruction [43].

Previously, prior biological knowledge has been incorporated into the neighbor selection method [44]. It relies on the Bayesian interpretation of LASSO and assigns two different prior distributions for connections that are present in the database and those are not. Recently, weighted graphical LASSO has been proposed to incorporate prior biological knowledge into graphical LASSO by assigning different weights to the entries of precision matrix [45]. We extend the original weighted graphical LASSO (wgLASSO) algorithm, explain this idea from a Bayesian perspective, and perform comprehensive comparisons between the proposed weighted graphical LASSO and competing data-driven network models (e.g., neighbor selection, graphical LASSO).

In the following, we introduce the extended weighted graphical LASSO algorithm wgLASSO in Section 3.2. Section 3.3 presents the results of wgLASSO based on the simulation data. Finally, Section 3.4 summarizes our work.

3.2 Weighted graphical LASSO

Consider a centered and scaled data matrix $\mathbf{X}_{n\times p}$ (i.e., $\sum_{i=1}^{n} x_{ij} = 0 , \sum_{i=1}^{n} x_{ij}^2 = 1$), it measures the intensities of p biomolecules on n samples, from a p-dimensional Gaussian distribution with zero means on each dimension and positive definite covariance matrix $\mathbf{\Sigma}_{p\times p}$ (i.e., $\mathbf{X}\sim N(\mathbf{0},\mathbf{\Sigma})$). Suppose the sample size n is far less than the variable number p (i.e., $n \ << p$), then the MLE of the precision matrix (i.e., $\mathbf{\Theta}=\mathbf{\Sigma}^{-1}$) doesn't exist since the sample covariance matrix (i.e., $\mathbf{S}$) is rank deficient. If we further assume $\mathbf{\Theta}$ is sparse, then a ℓ_1 regularization term can be added to the negative log-likelihood function $f(\mathbf{X}|\mathbf{\Theta}) = -\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S\Theta})$ for a sparse precision matrix estimation as shown in Equation 3.1. Graphical LASSO is an algorithm to efficiently solve Equation 3.1 by using block coordinate descent [34, 42]. Once $\widehat{\mathbf{\Theta}}$ is obtained, a non-zero element in $\widehat{\mathbf{\Theta}}$ (i.e., $\hat{\theta}_{jk} \neq 0$) indicates a conditional dependence between $\mathbf{x}_j$ and $\mathbf{x}_k$ given the rest. For network $\mathcal{G}=\{(j,k); 1\leq j<k\leq p\}$, we have $\hat{\mathcal{G}} = \left\{(j,k)\colon \hat{\theta}_{jk}\neq 0\right\}$.

$$\arg\min_{\mathbf{\Theta}\succ\mathbf{0}} -\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S\Theta}) + \lambda\|\mathbf{\Theta}\|_1 \tag{3.1}$$

where $\mathbf{\Theta}$ is the precision matrix, $\mathbf{\Theta}\succ\mathbf{0}$ is the constraint that $\mathbf{\Theta}$ has to be positive definite, $\mathbf{S}$ is the sample covariance matrix, tr denotes the trace, the sum of the diagonal elements in a matrix, $\|\mathbf{\Theta}\|_1$ represents the ℓ_1 norm, the sum of the absolute values of all the elements in $\mathbf{\Theta}$, and λ is the tuning parameter controlling the ℓ_1 shrinkage.

LASSO based estimates also have a Bayesian interpretation [46]. $\widehat{\mathbf{\Theta}}$ is the maximum a posteriori (MAP) estimate for the posterior distribution $p(\mathbf{\Theta}|\mathbf{X})$ with a Laplacian prior distribution $p(\mathbf{\Theta})$ as shown in Equation 3.2. The LASSO term $\lambda\|\mathbf{\Theta}\|_1$ in Equation 3.1 is now part of $p(\mathbf{\Theta}) = \exp\{-\lambda\|\mathbf{\Theta}\|_1\}$ with zero means and a scaling parameter λ. From a Bayesian perspective, $p(\mathbf{\Theta})$ encodes the prior biological knowledge of the network topology. For a database that contains only binary information (connecting or not) for a given gene pair, a natural way is to assign two different scaling parameters λ_1 and λ_2 for connecting pairs and those that are not connected, as shown in Equation 3.3. For connecting pairs, their Laplacian prior distribution is diffused, while that for non-connecting pairs is concentrated (i.e., $\lambda_1 \gg \lambda_2$). In another word, a larger penalty will be assigned to non-connecting pairs to increase the chance of their corresponding entries in $\mathbf{\Theta}$ to shrink to zero. In reality, tuning λ_1 and λ_2 at the same time involves two dimensional search, which is quite time-consuming for high-dimensional data. An extreme solution to set $\lambda_2 = 0$ links all the connecting gene pairs from the database in the graph, neglecting the fact that the database might contain some spurious connections for the disease under study.

$$p(\mathbf{\Theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathbf{X})} \propto \exp\{\log\det\mathbf{\Theta} - \mathrm{tr}(\mathbf{S\Theta})\} \times \exp\{-\lambda\|\mathbf{\Theta}\|_1\} \qquad (3.2)$$

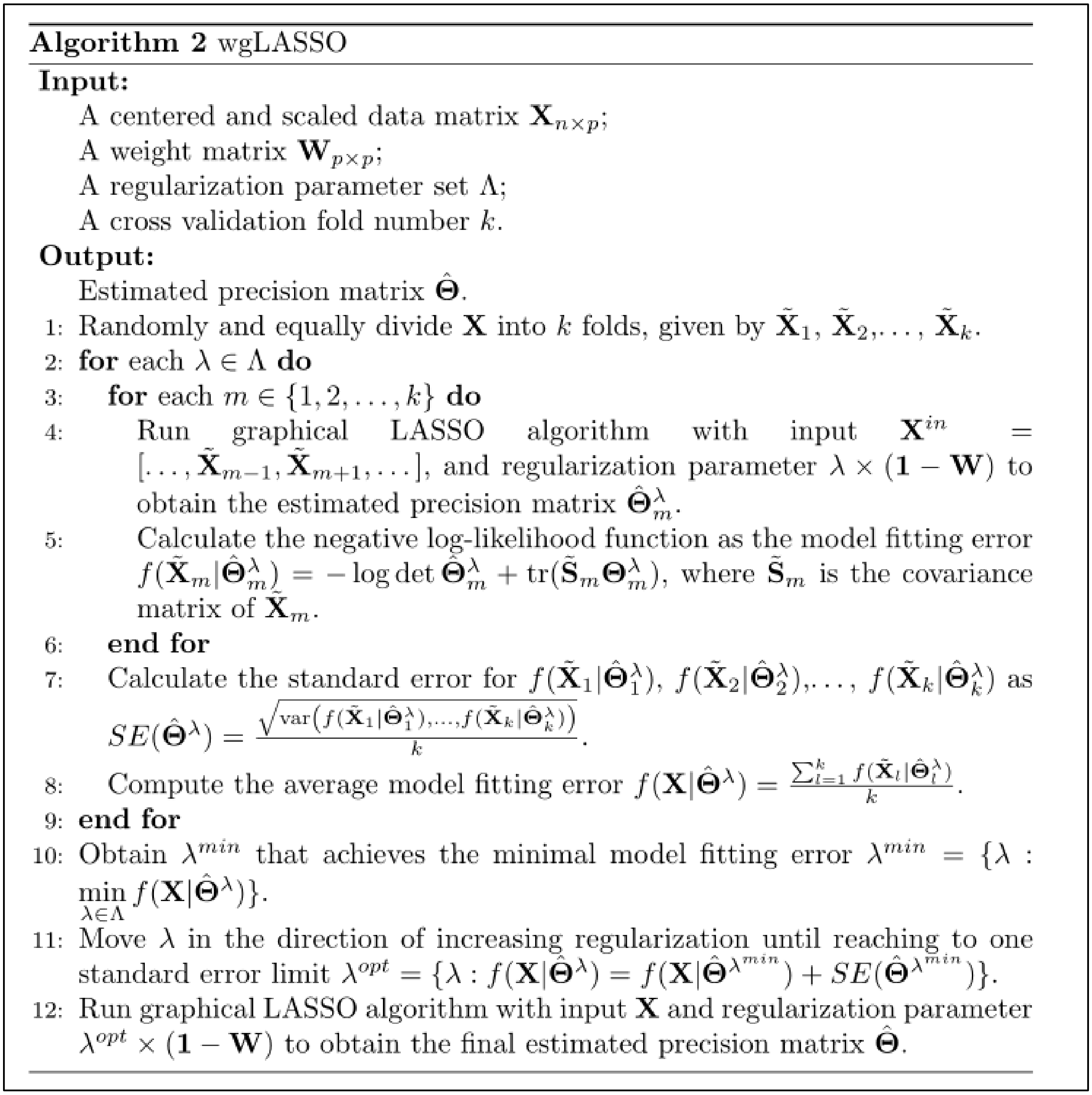
$$p(\mathbf{\Theta}) = \exp\{-\lambda_1 \sum \|\mathbf{\Theta}_{\text{non-connect}}\|_1 - \lambda_2 \sum \|\mathbf{\Theta}_{\text{connect}}\|_1\} \qquad (3.3)$$

Instead of using the binary information, a continuous confidence score is more suitable to incorporate prior biological knowledge into graphical LASSO. The confidence score can be obtained from multiple resources. For example, the GO semantic similarity between genes can be calculated using tools like GOSemSim or GOssTO [47, 48]. Additionally, an estimated functional association score for PPIs is provided by STRING database. We linearly scale this confidence score into the range [0,1] and create a weight matrix $\mathbf{W}_{\mathrm{p}\times\mathrm{p}}$. In $\mathbf{W}$, 1 indicates a complete trust for a gene pair to be connecting, 0 represents that no evidence supports a gene pair to be connecting. In this way, we can assign different penalties to various gene pairs as shown in Equation 3.4. Compared to Equation 3.3, Equation 3.4 also gives larger penalty for less likely connecting gene pair, but now there exists only one tuning parameter λ. For a fixed λ, R package glasso can solve Equation 3.4 efficiently given $\mathbf{W}$.

$$\arg\min_{\mathbf{\Theta}\succ\mathbf{0}} -\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S\Theta}) + \lambda\|(\mathbf{1} - \mathbf{W}) * \mathbf{\Theta}\|_1 \qquad (3.4)$$

where $\mathbf{1}$ is all 1 matrix, $\mathbf{W}$ is the weight matrix containing the confidence score for each gene pair and $*$ represents the component-wise multiplication.

For LASSO based optimization problem as shown in Equation 3.4, tuning the parameter λ is crucial since it controls the sparsity of the output $\widehat{\boldsymbol{\Theta}}$. Typically, λ is tuned by cross-validation, Bayesian information criterion (BIC), or stability selection [49]. Considering that BIC often leads to data under-fitting (i.e., over-sparse network) and stability selection requires extensive computation time for its internal sub-sampling procedure, we use cross validation with one standard error rule to select the optimal λ^{opt}. By using one standard error rule, we can achieve the simplest (most regularized) model whose error is within one standard deviation of the minimal error. Our weighted graphical LASSO algorithm runs as follows:



Algorithm 2 wgLASSO

```
Input:
    A centered and scaled data matrix $\mathbf{X}_{n\times p}$;
    A weight matrix $\mathbf{W}_{p\times p}$;
    A regularization parameter set $\Lambda$;
    A cross validation fold number $k$.
Output:
    Estimated precision matrix $\hat{\mathbf{\Theta}}$.
Randomly and equally divide $\mathbf{X}$ into $k$ folds, given by $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2$,..., $\tilde{\mathbf{X}}_k$.
for each $\lambda \in \Lambda$ do
    for each $m \in \{1, 2, \ldots, k\}$ do
        Run graphical LASSO algorithm with input $\mathbf{X}^{in} = [\ldots, \tilde{\mathbf{X}}_{m-1}, \tilde{\mathbf{X}}_{m+1}, \ldots]$, and regularization parameter $\lambda \times (\mathbf{1} - \mathbf{W})$ to obtain the estimated precision matrix $\hat{\mathbf{\Theta}}_m^\lambda$.
        Calculate the negative log-likelihood function as the model fitting error $f(\tilde{\mathbf{X}}_m|\hat{\mathbf{\Theta}}_m^\lambda) = -\log\det\hat{\mathbf{\Theta}}_m^\lambda + \mathrm{tr}(\tilde{\mathbf{S}}_m\mathbf{\Theta}_m^\lambda)$, where $\tilde{\mathbf{S}}_m$ is the covariance matrix of $\tilde{\mathbf{X}}_m$.
    end for
    Calculate the standard error for $f(\tilde{\mathbf{X}}_1|\hat{\mathbf{\Theta}}_1^\lambda)$, $f(\tilde{\mathbf{X}}_2|\hat{\mathbf{\Theta}}_2^\lambda)$,..., $f(\tilde{\mathbf{X}}_k|\hat{\mathbf{\Theta}}_k^\lambda)$ as $SE(\hat{\mathbf{\Theta}}^\lambda) = \frac{\sqrt{\mathrm{var}\left(f(\tilde{\mathbf{X}}_1|\hat{\mathbf{\Theta}}_1^\lambda),\ldots,f(\tilde{\mathbf{X}}_k|\hat{\mathbf{\Theta}}_k^\lambda)\right)}}{k}$.
    Compute the average model fitting error $f(\mathbf{X}|\hat{\mathbf{\Theta}}^\lambda) = \frac{\sum_{l=1}^k f(\tilde{\mathbf{X}}_l|\hat{\mathbf{\Theta}}_l^\lambda)}{k}$.
end for
Obtain $\lambda^{min}$ that achieves the minimal model fitting error $\lambda^{min} = \{\lambda : \min_{\lambda\in\Lambda} f(\mathbf{X}|\hat{\mathbf{\Theta}}^\lambda)\}$.
Move $\lambda$ in the direction of increasing regularization until reaching to one standard error limit $\lambda^{opt} = \{\lambda : f(\mathbf{X}|\hat{\mathbf{\Theta}}^\lambda) = f(\mathbf{X}|\hat{\mathbf{\Theta}}^{\lambda^{min}}) + SE(\hat{\mathbf{\Theta}}^{\lambda^{min}})\}$.
Run graphical LASSO algorithm with input $\mathbf{X}$ and regularization parameter $\lambda^{opt} \times (\mathbf{1} - \mathbf{W})$ to obtain the final estimated precision matrix $\hat{\mathbf{\Theta}}$.
```

3.3 Evaluation of dwgLASSO on simulated data

Biological networks are reported to be scale-free, that is the degree distribution of the network follows a power law [50]. We considered this scale-free property in generating simulation data using R package huge [51]. In huge, a scale-free network was built by inputting the node number p. The sparsity of the network s is fixed, depending on p. For example, when the node number is 100, the sparsity of the network is 0.02, meaning only 2% of all possible connections (i.e., $\frac{p\times(p-1)}{2}$) exist in the scale-free network. Once the scale-free network is built, huge creates the true precision matrix $\mathbf{\Theta}_{\mathrm{true}}$ based on the network topology and the positive definite constraint $\mathbf{\Theta}_{\mathrm{true}} \succ \mathbf{0}$ so that $\mathbf{\Sigma}_{\mathrm{true}} = (\mathbf{\Theta}_{\mathrm{true}})^{-1}$ exists. At last, simulation data $\mathbf{X}_{n\times p} \sim N(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{true}})$ was generated.

We generated simulation datasets with various p and n, as seen in Table 3.1. The weight matrix $\mathbf{W}$, which contains prior biological knowledge, was constructed based on $\mathbf{\Theta}_{\mathrm{true}}$. In reality, databases may also provide some spurious connections for the disease under study. To evaluate how the incorrect connections in $\mathbf{W}$ will impact weighted graphical LASSO, we introduced an additional metric acc. When $acc = 80\%$, we randomly reassigned 20% incorrect connections in $\mathbf{W}$. Specifically, $\mathbf{W}$ was created as follows. Initially, for zero entries in $\mathbf{\Theta}_{\mathrm{true}}$, the corresponding entries in $\mathbf{W}$ were also zero; for non-zero entries in $\mathbf{\Theta}_{\mathrm{true}}$, the corresponding entries in $\mathbf{W}$ were randomly generated from the uniform distribution $\mathcal{U}(0{,}1)$. Then, we randomly assigned incorrect connections into $\mathbf{W}$ based on the acc value while keeping the total connections in $\mathbf{W}$ the same as those in $\mathbf{\Theta}_{\mathrm{true}}$. Under the assumption that incorrect entries in $\mathbf{W}$ should have lower confidence scores compared to those of correct entries, we generated incorrect entries from the uniform distribution $\mathcal{U}(0{,}0.5)$.

We estimated the true network topology by using neighbor selection, graphical LASSO, and the proposed weighted graphical LASSO. For neighbor selection, two strategies were applied to deal with the inconsistency problem. Neighbor selection with “or” operator accepted inconsistent connections while neighbor selection with “and” operator rejected them. To make a fair comparison, we tuned the regularization parameter in each method to ensure the output network has the same sparsity as the true network (i.e., $s = 0.02$ for $p = 100$, $s = 0.004$ for $p = 500$). For the same n and p, we regenerated $\mathbf{X}_{n\times p}$ 100 times, calculated the false positives and false negatives of connections for each method, and listed the corresponding mean and standard deviation in Table 3.1. To evaluate how the incorrect connections in $\mathbf{W}$ would impact the performance of weighted graphical LASSO, we randomly reassigned 40% ($acc = 60\%$), 60% ($acc = 40\%$) and 80% ($acc = 20\%$) incorrect prior biological knowledge in $\mathbf{W}$. From Table 3.1, we can conclude that the estimated network from weighted graphical LASSO has significant lower false positives and false negatives, compared with those from neighbor selection and graphical LASSO in general. A decrease of acc in $\mathbf{W}$ would lead to more false positives and false negatives, but weighted graphical LASSO still outperforms neighbor selection and graphical LASSO methods even when the acc in $\mathbf{W}$ is only as moderate as 40%.

Table 3.1 The mean and standard deviation (in parenthesis) of false positives (FP) and false negatives (FN) for connections from neighbor selection (NS), graphical LASSO (gLASSO) and weighted graphical LASSO (wgLASSO) under different node number (p) and sample size (n) scenarios. The best performance is in bold.

p	n	NS 'or'		NS 'and'		gLASSO		wgLASSO ($acc = 60\%$)		wgLASSO ($acc = 40\%$)		wgLASSO ($acc = 20\%$)	
		FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
100	50	149.8 (16.8)	150.6 (10.4)	165.7 (15.0)	156.9 (10.3)	153.7 (23.1)	148.3 (10.8)	**111.8** **(17.1)**	**103.2** **(10.6)**	128.7 (17.6)	121.7 (10.9)	147.8 (21.6)	134.0 (10.4)
	100	113.3 (15.7)	110.6 (15.3)	131.6 (16.8)	122.2 (15.8)	113.8 (19.6)	112.0 (14.6)	**82.1** **(15.0)**	**74.4** **(12.5)**	92.7 (15.9)	86.9 (12.2)	107.4 (17.4)	99.3 (14.0)
	200	68.6 (13.4)	59.0 (18.2)	77.7 (14.8)	72.1 (20.8)	79.1 (17.0)	63.1 (18.9)	**51.4** **(11.3)**	**39.0** **(14.1)**	57.8 (13.4)	50.4 (15.4)	70.5 (15.8)	57.4 (16.3)
500	250	707.2 (41.6)	678.8 (77.0)	758.0 (42.8)	738.3 (81.7)	709.6 (47.6)	680.7 (77.0)	**480.1** **(35.8)**	**451.4** **(65.6)**	548.5 (39.3)	526.4 (60.4)	620.0 (42.8)	605.2 (76.9)
	500	425.4 (30.4)	453.4 (129.0)	473.2 (42.1)	492.7 (133.6)	430.9 (40.3)	468.4 (128.8)	**276.6** **(26.4)**	**290.3** **(87.4)**	329.7 (30.8)	313.0 (105.9)	377.9 (35.4)	395.3 (113.5)
	1000	175.3 (21.9)	163.6 (116.7)	189.2 (26.8)	176.6 (118.4)	198.6 (27.5)	185.6 (126.1)	**108.5** **(17.5)**	**110.2** **(75.5)**	130.1 (20.7)	134.5 (87.5)	159.8 (24.9)	151.4 (96.2)

To make more comprehensive comparison, we plotted precision recall curve to evaluate the performance of neighbor selection, graphical LASSO and weighted graphical LASSO. For a sparse network, precision recall curve is a better visualization plot to evaluate the performance of different methods than ROC curve. We selected regularization parameter ranging from large to small in log-scale, ran the above methods with $p = 100$, $n = 50$ and $acc = 40\%$ in $\mathbf{W}$, computed the precision and recall, and generated the plot as shown in Figure 3.1. From Figure 3.1, weighted graphical LASSO displays a clear improvement over the neighbor selection and graphical LASSO methods. This agrees with our expectation since weighted graphical LASSO considers whether the connection has supporting evidence from database and how good it fits the data simultaneously.

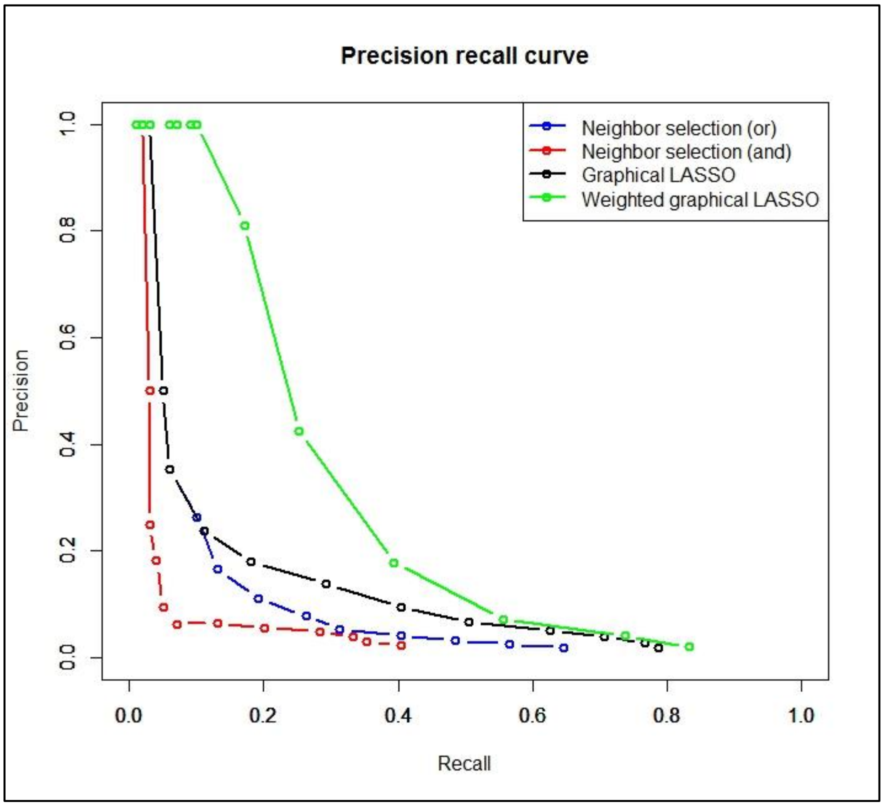


Figure 3.1, Precision recall curves for neighbor selection, graphical LASSO and weighted graphical LASSO under $p = 100$, $n = 50$ and $acc = 40\%$.

3.4 Conclusion

In Chapter 3, we extend a novel network reconstruction method, wgLASSO to integrate prior biological knowledge into data-driven model (e.g., graphical LASSO). Simulation results show that weighted graphical LASSO can achieve better performance in building biologically relevant networks than purely data-driven models (e.g., neighbor selection and graphical LASSO) even when a moderate level of information is available as prior biological knowledge. We also develop an open source R package to share wgLASSO with the scientific community at Github (https://github.com/Hurricaner1989/dwgLASSO-R-package).

4 Differential network analysis using dwgLASSO

4.1 Introduction

Recent advances in high-throughput technique enables the generation of a large amount of omic data at different levels such as genomics, transcriptomics, proteomics metabolomics, etc. Typically, a differential expression analysis (e.g., student's *t*-test, SAM, Empirical Bayes, etc.) is performed to identify biomolecules with significant changes between biologically disparate groups [52-54]. However, independent studies for the same clinical types of patients often lead to different sets of significant biomolecules and had only few in common [1]. This may be attributed to the fact that biomolecules are members of strongly intertwined biological pathways and are highly interactive with each other. Without considering these interactions, differential expression analysis will easily yield biased result and lead to a fragmented picture.

Network-based methods provide a natural framework to study the interactions among biomolecules [3]. This includes purely data-driven methods such as relevance network, Bayesian network, GGM introduced in Chapter 2. Additionally, other data resources such as literature, functional annotation, biomolecular interaction and sequence information are also commonly used [55, 56]. Our proposed wgLASSO algorithm in Chapter 3 used PPI information as an extra data resource. Once the networks are reconstructed, there exist multiple tools to prioritize the biomolecules selected from differential expression analysis based on the topology of the network [57-59]. Unlike these conventional network-based differential expression analysis methods, we would like to build group specific network and explore the topological changes between biological disparate groups, which has been reported to lead to new discoveries that cannot be identified by typical differential expression analysis [60-62]. For example, high-degree nodes (i.e., hubs) that only exist in one of the biologically disparate groups may indicate the regulatory rule of those hub biomolecules only in that group. Knowledge-fused differential dependency network (KDDN) is a recently proposed method to construct knowledge incorporated network that can show the rewiring connections between two groups [62]. An open-source Cytoscape app is available for easy implementation [63].

Considering this, we propose a new algorithm (differentially weighted graphical LASSO, dwgLASSO) for differential network analysis. This is achieved by building separate networks for biologically disparate groups using wgLASSO introduced in Chapter 3, exploring the topological changes between different groups based on the node degrees, and prioritizing top ranking biomolecules from conventional differential expression analysis as shown in Figure 4.1. Other previously reported methods include those that focus on integrating prior biological knowledge into a data-driven network model to identify sub-networks that are related to the disease under study [64, 65]. Our work differs with these methods since we compute a differential network score for each biomolecule and prioritize them for subsequent analysis rather than generating a list of sub-networks for biological interpretation. Also, methods that directly incorporate biological networks or prior biological knowledge into statistical models for classification and regression tasks have been reported [66, 67]. The rationale is that functionally linked biomolecules tend to be co-regulated and co-expressed, and therefore should be treated similarly in the statistical model. Our work leaves the statistical model untouched. Instead, it focuses on using the best set of biomarker candidates as an input to the statistical model. This is considered to have advantages over providing multiple linked biomolecules from the network whose expression values have similar patterns. We show the application of dwgLASSO on two independent microarray datasets from breast cancer patients for survival time prediction, and on TCGA RNA-seq data acquired from patients with hepatocellular carcinoma (HCC) for classification task between tumor samples and their corresponding non-tumorous liver tissues [68, 69]. For microarray dataset, compared with the top 10 significant genes selected by conventional differential gene expression analysis method, the top 10 significant genes selected by dwgLASSO in the dataset from Bild *et al.* led to a significantly improved survival time prediction in the independent dataset from van de Vijver *et al.* Among the 10 genes selected by dwgLASSO, UBE2S, SALL2, XBP1 and KIAA0922 have been confirmed by literature survey to be highly relevant in breast cancer biomarker discovery study. For TCGA RNA-seq dataset, improved sensitivity, specificity and area under curve (AUC) were observed when comparing dwgLASSO with conventional differential gene expression analysis method.

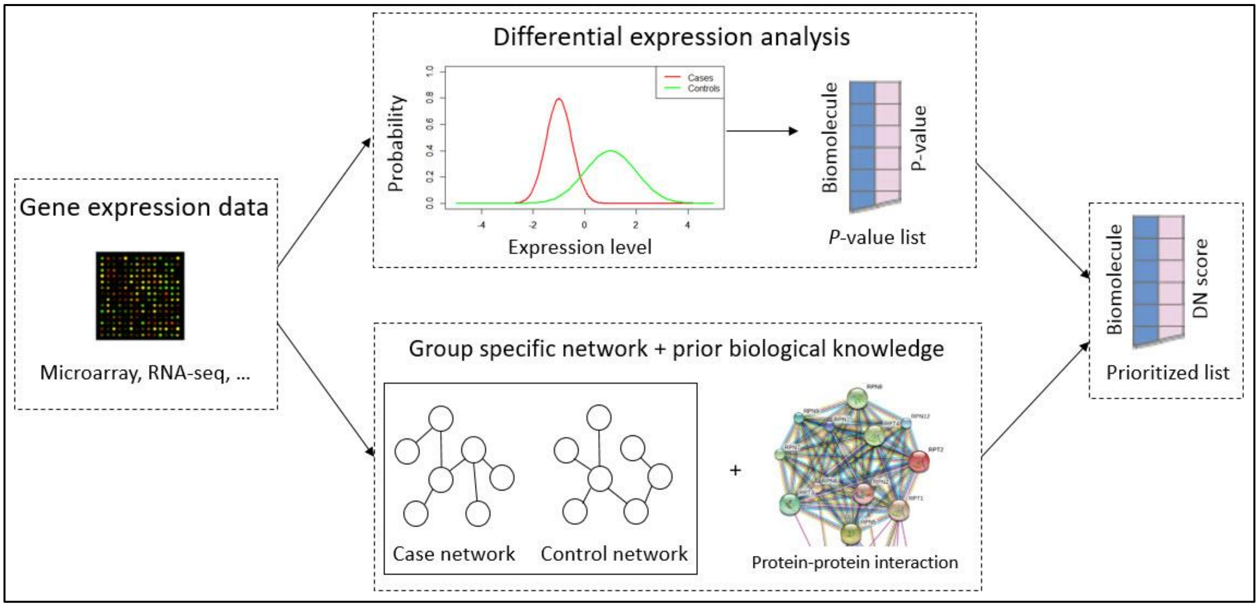


Figure 4.1 An overview of dwgLASSO. The input is gene expression data (e.g., Microarray, RNA-seq data, etc.) and the output is a prioritized list based on the differential network (DN) score defined within dwgLASSO.

In the following, we introduce the proposed dwgLASSO for differential network analysis in Section 4.2. Section 4.3 presents the results of dwgLASSO based on real microarray data and RNA-seq data. Finally, Section 4.4 summarizes our work.

4.2 Differentially weighted graphical LASSO

Figure 4.2 shows the framework of a novel algorithm (dwgLASSO) we developed for differential network analysis. dwgLASSO prioritizes the selected biomolecules from the conventional differential expression analysis based on the topological changes between the group-specific networks built by wgLASSO.

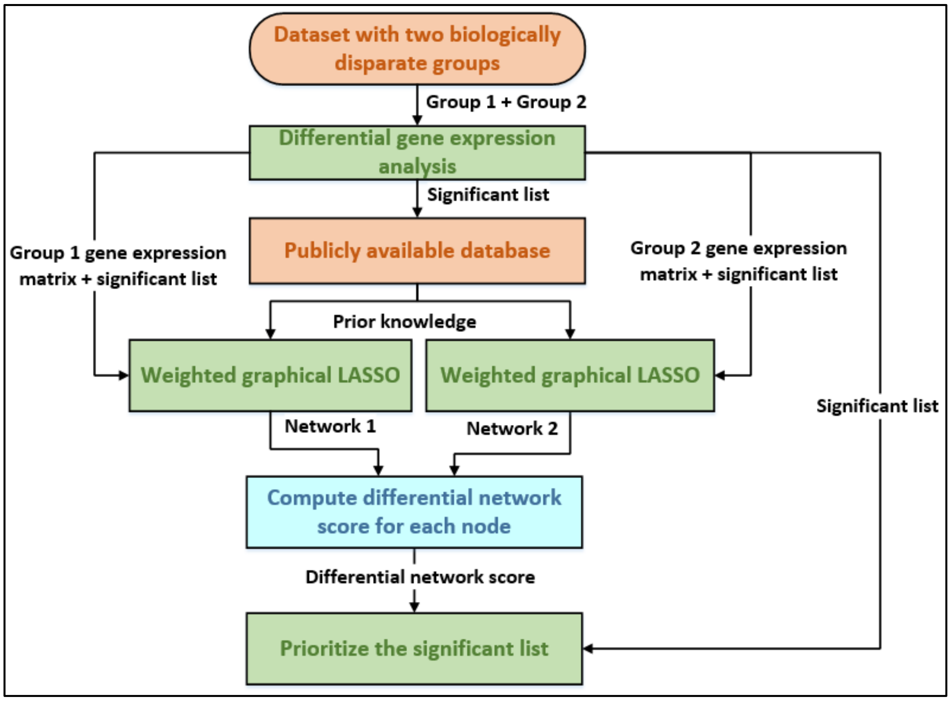
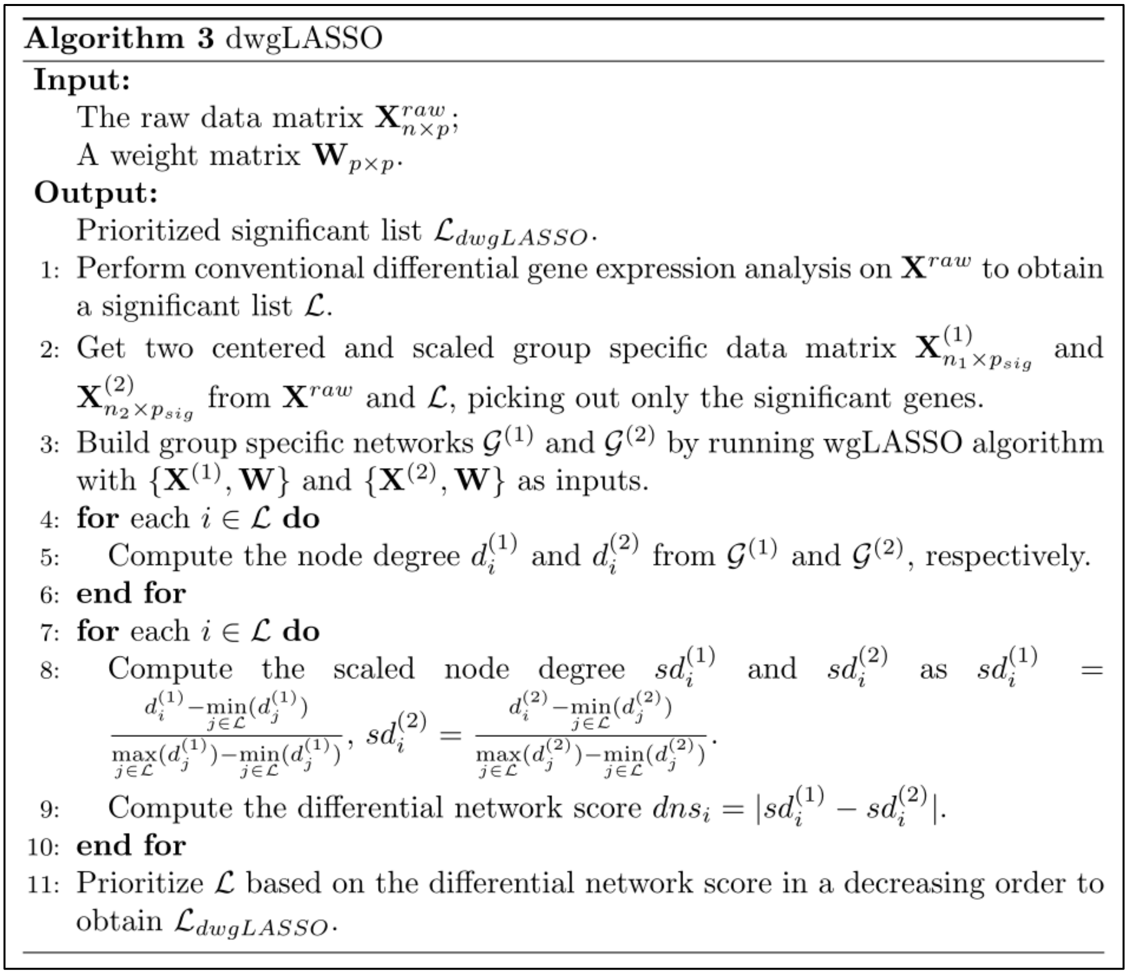

Dataset with two biologically disparate groups
Group 1 + Group 2
Differential gene expression analysis
Significant list
Publicly available database
Group 1 gene expression matrix + significant list
Group 2 gene expression matrix + significant list
Prior knowledge
Weighted graphical LASSO
Weighted graphical LASSO
Significant list
Network 1
Network 2
Compute differential network score for each node
Differential network score
Prioritize the significant list

Figure 4.2 Framework for dwgLASSO to perform network-based differential expression analysis.

Specifically, dwgLASSO first performs differential expression analysis to obtain a list of significant or top ranking biomolecules whose expression levels differ between the two biologically disparate groups. Then from these biomolecules, dwgLASSO builds group specific networks using wgLASSO. After the networks are constructed, dwgLASSO calculates a differential score for each biomolecule from last step based on the topological changes between the two groups. In calculating the differential score, dwgLASSO first computes the node degree for each biomolecule in both networks, meaning the number of neighbors each node is connected with. Then considering the size of the two group-specific networks are different, the node degrees are linearly scaled into a range $[0, 1]$ for fair comparisons between the two groups. At last, the differential score for one biomolecule is computed as the absolute value of the difference between the two associated scaled node degrees from each group. With the differential scores, dwgLASSO prioritizes the selected biomolecules from differential expression analysis in a decreasing order. We believe the prioritized list can achieve better prediction performance since dwgLASSO integrates information at the biomolecular expression and network structure levels. More than that, the incorporation of prior biological knowledge in building group-specific network (i.e., wgLASSO) is more likely to identify biologically meaningful biomolecules. Detailed algorithm for dwgLASSO is shown below:



Algorithm 3 dwgLASSO

```
Input:
    The raw data matrix $\mathbf{X}^{raw}_{n \times p}$;
    A weight matrix $\mathbf{W}_{p \times p}$.
Output:
    Prioritized significant list $\mathcal{L}_{dwgLASSO}$.
Perform conventional differential gene expression analysis on $\mathbf{X}^{raw}$ to obtain a significant list $\mathcal{L}$.
Get two centered and scaled group specific data matrix $\mathbf{X}^{(1)}_{n_1 \times p_{sig}}$ and $\mathbf{X}^{(2)}_{n_2 \times p_{sig}}$ from $\mathbf{X}^{raw}$ and $\mathcal{L}$, picking out only the significant genes.
Build group specific networks $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ by running wgLASSO algorithm with $\{\mathbf{X}^{(1)}, \mathbf{W}\}$ and $\{\mathbf{X}^{(2)}, \mathbf{W}\}$ as inputs.
for each $i \in \mathcal{L}$ do
    Compute the node degree $d_i^{(1)}$ and $d_i^{(2)}$ from $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, respectively.
end for
for each $i \in \mathcal{L}$ do
    Compute the scaled node degree $sd_i^{(1)}$ and $sd_i^{(2)}$ as $sd_i^{(1)} = \frac{d_i^{(1)} - \min_{j \in \mathcal{L}}(d_j^{(1)})}{\max_{j \in \mathcal{L}}(d_j^{(1)}) - \min_{j \in \mathcal{L}}(d_j^{(1)})}$, $sd_i^{(2)} = \frac{d_i^{(2)} - \min_{j \in \mathcal{L}}(d_j^{(2)})}{\max_{j \in \mathcal{L}}(d_j^{(2)}) - \min_{j \in \mathcal{L}}(d_j^{(2)})}$.
    Compute the differential network score $dns_i = |sd_i^{(1)} - sd_i^{(2)}|$.
end for
Prioritize $\mathcal{L}$ based on the differential network score in a decreasing order to obtain $\mathcal{L}_{dwgLASSO}$.
```

4.3 Evaluation of dwgLASSO on transcriptomic data

4.3.1 dwgLASSO on microarray data

We applied the proposed network-based differential expression analysis algorithm dwgLASSO on two breast cancer microarray datasets: Bild *et al.*'s and van de Vijver *et al.*'s datasets [68, 69]. The former includes 158 patients with all their survival records, and is used for training. We excluded patients with less than 5-year follow-up time. Among the remaining patients, 42 with less than 5-year survival during the follow-up time are considered high risk group while the other 60 form the low risk group. van de Vijver *et al.*'s dataset contains 295 breast cancer patients, together with their survival records, and is used for independent testing. Both datasets are available at PRECOG website (https://precog.stanford.edu/), an online repository for querying cancer gene expression and clinical data, and have been preprocessed for subsequent statistical analysis [70]. The raw Bild *et al.*'s and van de Vijver *et al.*'s datasets are also available at Gene Expression Omnibus (GSE3143) and R package seventyGeneData, respectively.

Our interest is to obtain a prioritized gene list based on dwgLASSO for more accurate survival time prediction. The workflow is shown in Figure 4.3. We first performed univariate analysis on Bild *et al.*'s dataset to select a list of statistically significant genes based on concordance index between the expression value and survival time [71]. This lead to a total of 58 genes whose adjusted p-values were less than 0.05. The inflation of Type I error caused by multiple testing was controlled by the false discovery rate (FDR) using the Benjamini-Hochberg procedure. The total 58 significant genes are included in Table S1 of the Supplementary Materials along with their associated adjusted p-values [72]. We then applied wgLASSO algorithm to build two separate networks using the total 58 significant genes for the high risk and low risk groups, respectively. The weight matrix $\mathbf{W}$ was constructed based on the confidence scores outputted from STRING database after inputting the 58 significant genes to investigate the PPIs among them. For gene pairs with no confidence scores from STRING, we assigned the corresponding entries in $\mathbf{W}$ to zeros. In wgLASSO, we performed 10-fold cross validation and chose the optimal tuning parameter λ^{opt} by one standard error rule. Figure 4.4 shows our chose of λ^{opt}: $\lambda^{opt} = 0.223$ for high risk group and $\lambda^{opt} = 0.184$ for low risk group, respectively. From the networks, we calculated the node degree for each gene in two groups (d_i^h, d_i^l), scale them based on the network size ($ds_{\mathrm{i}}^h, ds_{\mathrm{i}}^l$), and compute the differential network score ($dns_i =$

$|ds_{\mathrm{i}}^{h}-ds_{\mathrm{i}}^{l}|$). At last, we prioritized the 58 significant genes based on the differential scores in a decreasing order.

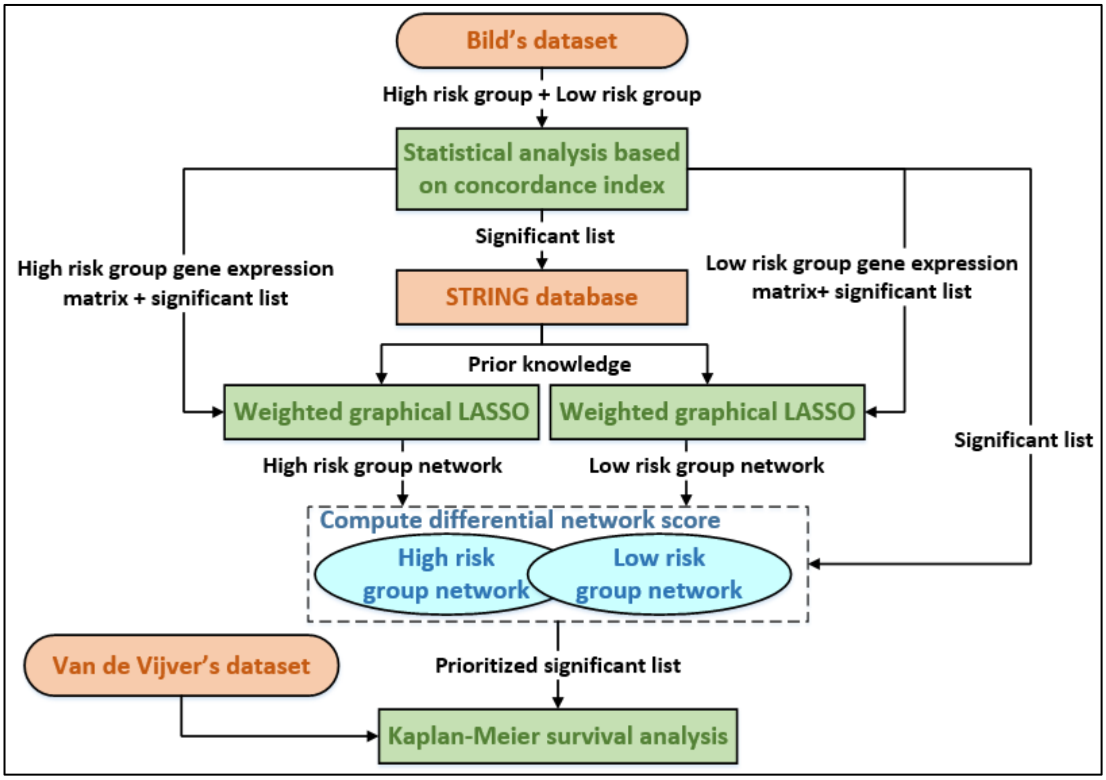

Bild's dataset
High risk group + Low risk group
Statistical analysis based on concordance index
Significant list
STRING database
High risk group gene expression matrix + significant list
Low risk group gene expression matrix+ significant list
Prior knowledge
Weighted graphical LASSO
Weighted graphical LASSO
Significant list
High risk group network
Low risk group network
Compute differential network score
High risk group network
Low risk group network
Van de Vijver's dataset
Prioritized significant list
Kaplan-Meier survival analysis

Figure 4.3 Workflow to obtain prioritized gene list based on dwgLASSO for more accurate survival time prediction.

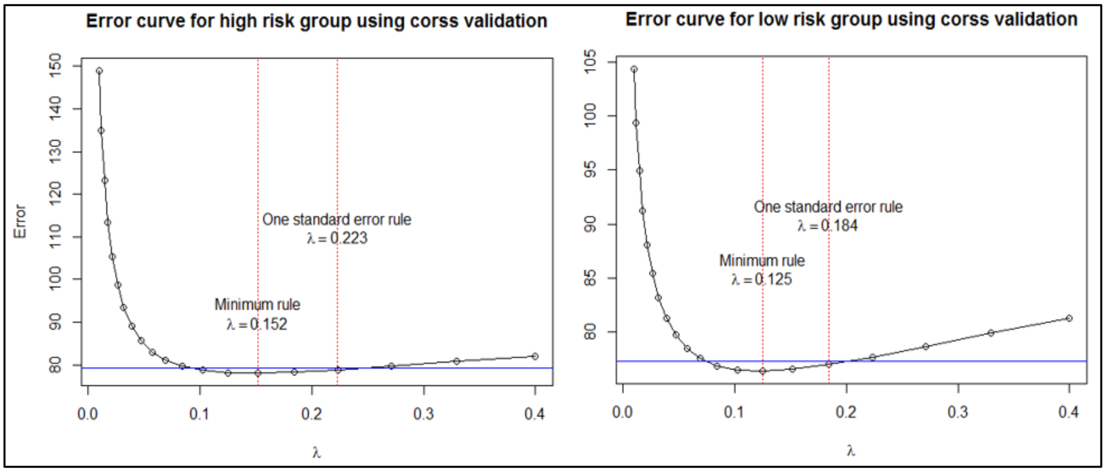


Figure 4.4 Error curves to choose optimal tuning parameter λ^{opt} using 10 fold cross validation by one standard error rule for high risk and low risk groups. The blue line indicates the one standard error for λ^{min} in the direction of increasing regularization.

To evaluate whether dwgLASSO could lead to more accurate survival time prediction, we tested the prioritized gene list using different methods on the independent van de Vijver *et al.*'s dataset. The 295 patients were divided into high risk and low risk groups according to the risk scores calculated using multivariate Cox regression from the top 10 genes based on dwgLASSO, KDDN, and conventional differential gene expression analysis (i.e., concordance index). Unlike dwgLASSO that builds group-specific networks, KDDN generates only one network with all rewiring connections. From the network constructed by KDDN, we computed the node degree for each gene to help prioritize the gene list. Kaplan-Meier survival analysis was then conducted to evaluate the performance of the above three scenarios, resulting to the survival curves in Figure 4.5A, 4.5B and 4.5D. To evaluate how much the incorporation of prior biological knowledge contributes to achieving the better prediction in dwgLASSO, we tested an additional top 10 ranking genes based on dwgLASSO with no prior biological knowledge incorporated (i.e., $\mathbf{W} = \mathbf{0}$). The survival curve is shown in Figure 4.5C. As expected, dwgLASSO with no prior biological knowledge incorporated is equivalent to graphical LASSO in building group-specific networks (Figure 4.2). From Figure 4.5, the top 10 genes from dwgLASSO with prior biological knowledge incorporated yielded the best performance (p-value=7.01e^{-7}, hazard ratio=3.325), compared to the top 10 genes from KDDN (p-value=7.46e^{-7}, hazard ratio=3.304), the top 10 genes based on dwgLASSO with no prior biological knowledge incorporated (p-value=0.00031, hazard ratio=2.316), and the top 10 genes based on concordance index (p-value=0.002, hazard ratio=2.037). We believe the improved performance achieved by dwgLASSO and KDDN are due to the extra information provided from the topological changes between high risk and low risk groups. Also, dwgLASSO and KDDN benefit from incorporating prior biological knowledge to obtain more reliable and biologically relevant genes shared across independent datasets, leading to better prediction performance than those that do not use prior biological knowledge (Figure 4.5). Table 4.1 presents the top 10 genes selected based on concordance index and dwgLASSO with prior biological knowledge incorporated. The top 10 genes from the other methods are presented in Supplementary Materials Table S2 [72].

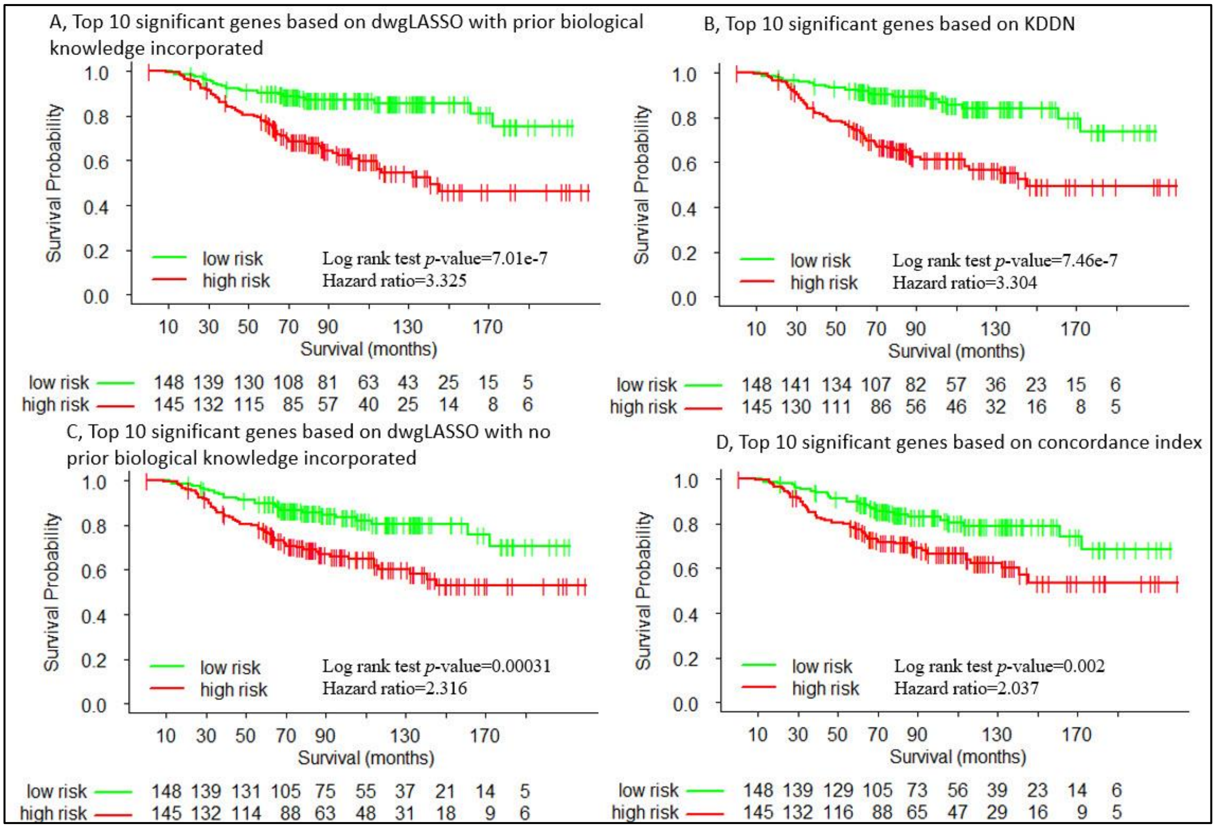


Figure 4.5 Survival curves for A) top 10 ranking genes based on dwgLASSO with prior knowledge incorporated, B) top 10 ranking genes based on KDDN, C) top 10 ranking genes based on dwgLASSO with no prior knowledge incorporated, D) top 10 ranking genes based on concordance index.

Among the top 10 ranking genes based on dwgLASSO in Table 4.1, UBE2S has been reported to be over-expressed in breast cancer [73]. The authors showed UBE2S knockdown suppressed the malignant characteristics of breast cancer cells, such as migration, invasion, and anchorage-independent growth. SALL2 has also been reported as a predictor of lymph node metastasis in breast cancer [74]. Unlike UBE2S, SALL2 was identified as a tumor suppressor gene that can suppress cell growth when over-expressed [75]. Additionally, XBP1 has been reported to be activated in triple-negative breast cancer and has a pivotal role in the tumorigenicity and progression of this breast cancer subtype [76]. KIAA0922 has also been reported as a novel inhibitor of Wnt signaling pathway, which is closely related to breast cancer [77]. None of UBE2S, SALL2, XBP1 and KIAA0922 is among the top 10 significant genes based on concordance index according to Table 4.1.

Table 4.1 The top 10 ranking genes based on concordance index and dwgLASSO with prior biological knowledge incorporated, along with their adjusted *p*-value. Common genes are in bold.

Top 10 genes based on concordance index		Top 10 genes based on weighted graphical LASSO	
Gene symbol	Adjusted *p*-value	Gene symbol	Adjusted *p*-value
BTD	0.000167029	SALL2	0.018149333
FKTN	0.000424976	UBE2S	0.015577505
LRRC17	0.000424976	**RAB11FIP5**	0.001638818
RAB11FIP5	0.001638818	KIAA1467	0.005012636
EMX2	0.002384716	XBP1	0.005019825
HNRNPAB	0.002384716	KIAA0922	0.021163875
TKT	0.002805234	**EMX2**	0.002384716
LANCL1	0.003481701	OAZ2	0.040090787
TFF3	0.003481701	NDC80	0.030630047
USF2	0.004094746	CCT5	0.048116117

In Figure 4.6, we showed the neighbors of UBE2S and SALL2 in the high risk and low risk groups based on the networks created by wgLASSO from Bild *et al.*'s dataset. UBE2S is over-expressed in the high risk group while SALL2 is under-expressed. This agrees with that UBE2S is a promoting breast cancer gene while SALL2 is a suppressor breast cancer gene. Additionally, UBE2S has higher scaled node degree in the high risk group while SALL2 has higher scaled node degree in the low risk group ($ds^{h}_{\text{UBE2S}} = 0.286$, $ds^{l}_{\text{UBE2S}} = 0.778$, $ds^{h}_{\text{SALL2}} = 1.0$, $ds^{l}_{\text{SALL2}} = 0.444$). This shows, as a promoting breast cancer gene, UBE2S is more actively connected with its neighbors in the high risk group while, the suppressor breast cancer gene, SALL2 is more actively connected with its neighbors in the low risk group.

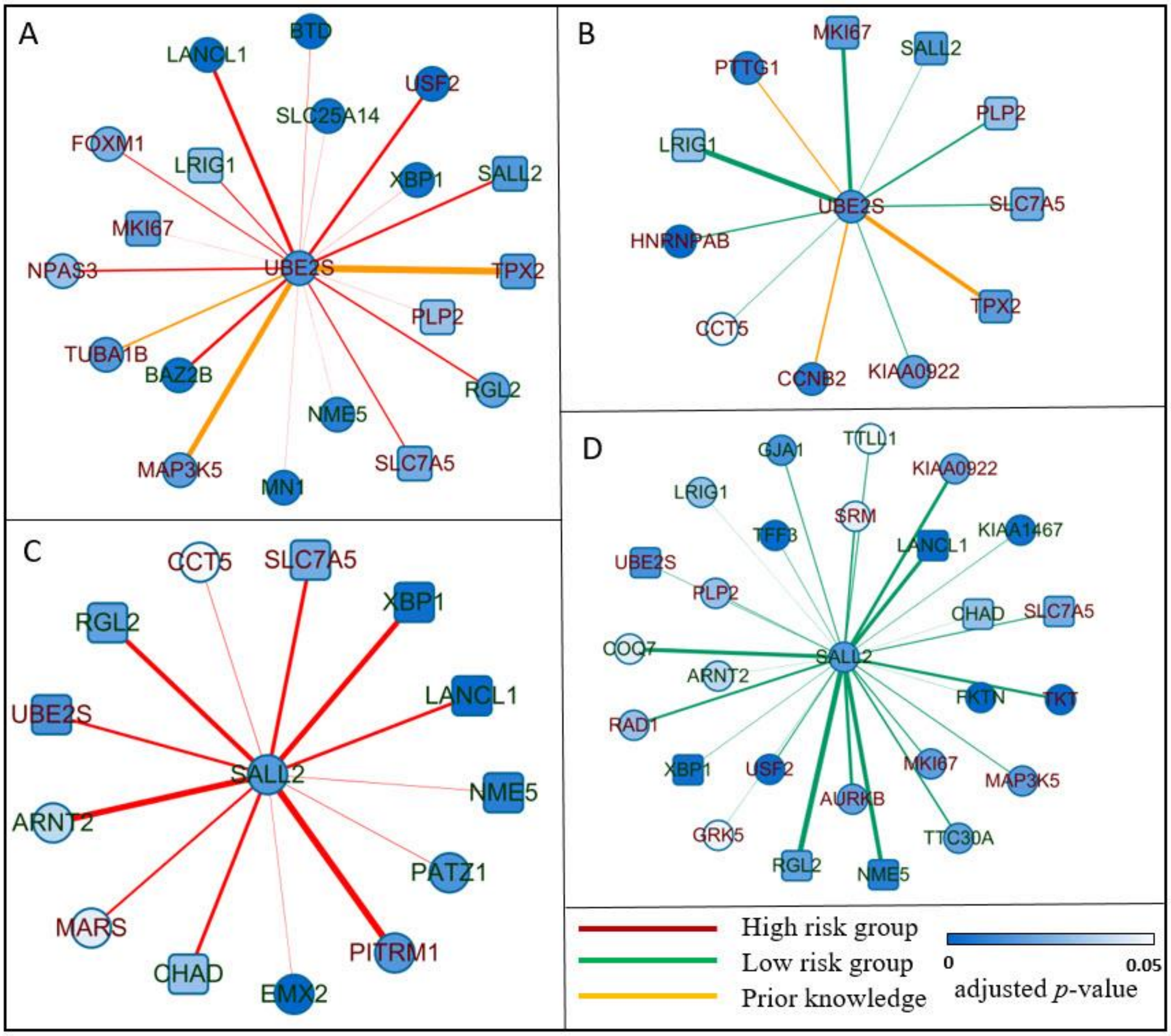


Figure 4.6 A) neighbors of UBE2S in the high risk group, B) neighbors of UBE2S in the low risk group, C) neighbors of SALL2 in the high risk group, D) neighbors of SALL2 in the low risk group. Label colors represent over- (red) or under- (green) expression in the high risk group. Node shapes indicate unique (circle) or shared (rectangle) genes between the two groups. Node colors show the significance of the

gene expression between the two groups. Yellow edges represent interactions recorded in the STRING database. Thickness of the edge indicates the strength of the interaction.

In Figure 4.6, yellow edges represent connections that have been supported from STRING database. We can see that this kind of connections based on prior biological knowledge are not always showing up from our wgLASSO algorithm. This is a nice property since prior biological knowledge only provides evidence. We still need the support from the data to make a connection. Therefore, by integrating prior biological knowledge into data-driven models, we expect to build more robust and biologically relevant networks. Table 4.2 shows the survival time prediction performance when the top 5, top 10 and top 15 significant genes are selected by each of the four methods as input to the multivariate Cox regression model (Figure 4.3). In all three cases, the proposed dwgLASSO algorithm with prior biological knowledge incorporated achieved the best performance, followed by KDDN and dwgLASSO without prior biological knowledge incorporated. The method that relies purely on concordance index had the least performance.

Table 4.2 The survival time prediction performance (p-value and hazard ratio) for the top 5, top 10 and top 15 significant genes based four different methods: (1) concordance index, DEA; (2) dwgLASSO with no prior biological knowledge incorporated, dwgLASSO (no prior); (3) KDDN, and (4) dwgLASSO with prior biological knowledge incorporated, dwgLASSO (prior). The best performance is marked in bold when the gene number is fixed.

Methods	Top 5 significant genes		Top 10 significant genes		Top 15 significant genes	
	p-value	hazard ratio	p-value	hazard ratio	p-value	hazard ratio
DEA	0.0073	1.851	2.00E-03	2.037	4.00E-04	2.274
dwgLASSO (no prior)	0.0066	1.864	3.10E-04	2.316	4.60E-06	2.969
KDDN	0.0022	2.028	7.46E-07	3.304	8.04E-06	2.889
dwgLASSO (prior)	**0.0013**	**2.104**	**7.01E-07**	**3.325**	**9.37E-07**	**3.25**

4.3.2 dwgLASSO on RNA-seq data

Using UCSC Cancer Genomics Browser, we obtained TCGA RNA-seq data (level 3) acquired from patients with HCC [78]. The RNA-seq data was acquired by analysis of 423 liver tissues, including 371 primary tumor, 50 solid normal, and 2 recurrent tumor samples based on Illumina HiSeq 2000 RNA Sequencing platform and mapped onto the human genome coordinates using UCSC cgData HUGO probeMap. Among the 371 primary tumor samples, 50 of them can find its corresponding solid normal samples. To evaluate dwgLASSO on RNA-seq data, we apply a workflow shown in Figure 4.7. We first picked out the 100 samples whose tumor tissues and their corresponding non-tumorous tissues can both be found. Randomly, we selected 60 of them (30 tumor samples and their corresponding normal samples) as the training dataset. The remaining 40 samples (20 tumor samples and their corresponding normal samples) were used as testing dataset 1. Considering testing dataset 1 only contains 40 samples, we created testing dataset 2 by combining the above 40 samples and the remaining 321 tumor samples whose corresponding normal samples cannot be found. With testing datasets 1 and 2, we evaluated the performance of dwgLASSO on both balanced and large sample size datasets. Specifically, we preprocessed RNA-seq data using R package DESeq2 on the training dataset [79]. From DESeq2, we selected statistically significant genes whose adjusted p-values were less than 0.01 for subsequent analysis. At this step, the number of significant genes is typically between 1000 and 2000. We prioritized the significant gene list based on dwgLASSO. From the prioritized gene list, the top 5 genes were selected to train a logistic regression classifier to distinguish tumor and normal samples. The trained logistic regression classifier was finally evaluated on testing datasets 1 and 2. To compare dwgLASSO with other methods, we also prioritized the significant gene list based on adjusted p-value from DESeq2, dwgLASSO without prior biological knowledge incorporated and KDDN, built logistic regression classifier using the top 5 genes on the prioritized list and evaluated the trained classifier on the testing datasets 1 and 2. In addition to the comparison of dwgLASSO with conventional differential gene expression analysis and network-based method without prior biological knowledge incorporated, we further compared dwgLASSO with a competing network-based method with prior biological knowledge incorporated (i.e., KDDN).

The above procedure was repeated 100 times and the means and standard deviations for sensitivity, specificity and area under curve (AUC) were calculated using testing datasets 1 and 2 as shown in Table 4.3. In agreement with microarray data, network-based methods with prior biological knowledge incorporated yielded the best performance, followed by network-based method without prior biological knowledge incorporated, and the conventional gene expression analysis method was the worst. This is expected since both dwgLASSO and KDDN methods take into account the changes of genes at gene expression and network topology levels, and incorporate prior biological knowledge into their network models.

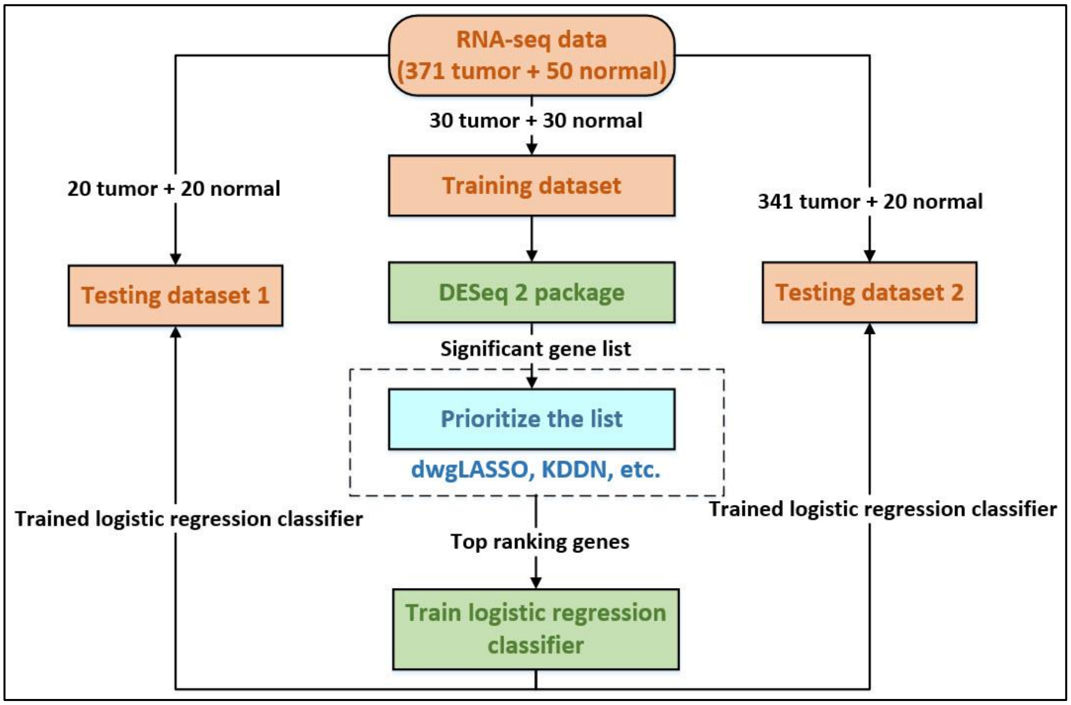


Figure 4.7 Workflow of dwgLASSO for more accurate classification prediction on RNA-seq data.

Table 4.3 The mean and standard deviation (in parenthesis) of sensitivity, specificity and area under curve (AUC) calculated for four methods: (1) conventional differential gene expression analysis, DEA; (2) dwgLASSO with no prior biological knowledge incorporated, dwgLASSO (no prior); (3) KDDN; and (4) dwgLASSO with prior biological knowledge incorporated, dwgLASSO (prior). The best performance is marked in bold.

Methods	Testing dataset 1			Testing dataset 2		
	specificity	sensitivity	AUC	specificity	sensitivity	AUC
DEA	0.95 (0.07)	0.913 (0.06)	0.951 (0.04)	0.950 (0.07)	0.941 (0.04)	0.983 (0.01)
dwgLASSO (no prior)	**0.988 (0.03)**	0.888 (0.11)	0.972 (0.02)	**0.988 (0.03)**	0.956 (0.05)	0.990 (0.01)
KDDN	0.963 (0.08)	**0.950 (0.04)**	0.980 (0.02)	0.963 (0.08)	0.939 (0.03)	0.989 (0.01)
dwgLASSO (prior)	**0.988 (0.03)**	0.950 (0.07)	**0.982 (0.03)**	**0.988 (0.03)**	**0.965 (0.03)**	**0.994 (0.01)**

4.4 Conclusion

In Chapter 4, we propose a novel differential network analysis algorithm dwgLASSO for better identification of biomolecules associated with biologically disparate groups. We demonstrate the performance of dwgLASSO in survival time prediction using two independent microarray breast cancer datasets previously published by Bild *et al.* and van de Vijver *et al.* The top 10 ranking genes selected by dwgLASSO based on the dataset from Bild *et al.* lead to a significantly improved survival prediction on the dataset from van de Vijver *et al.*, compared with the top 10 significant genes obtained by conventional differential gene expression analysis. Among the top 10 genes selected by dwgLASSO, UBE2S, SALL2, XBP1 and KIAA0922 have been previously reported to be relevant in breast cancer biomarker discovery study. We also test dwgLASSO using TCGA RNA-seq data acquired from patients with HCC on tumors samples and their corresponding non-tumorous liver tissues. Improved sensitivity, specificity and AUC are observed when comparing dwgLASSO with conventional differential gene expression analysis method. At last, we develop an open source R package to share dwgLASSO with the scientific community at Github (https://github.com/Hurricaner1989/dwgLASSO-R-package).

5 INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery

5.1 Introduction

In Chapter 2 and Chapter 3, we developed novel methods to reconstruct biological network using partial correlation and incorporate prior biological knowledge in building biologically more relevant network. In Chapter 4, we developed a novel method dwgLASSO to perform differential network analysis. There has been a growing interest in differential network (DN) analysis recently [5, 61, 62]. In a differential network, the connection represents a statistically significant change in the pairwise association between two biomolecules on distinct groups. Its goal is to identify sub-networks (i.e., connected biomolecules) that are dysfunctional in a given disease state. Compared to group-specific networks in dwgLASSO, differential network analysis only focuses on the connections that show significant changes between different groups. This is desirable since we are more interested in the difference between biologically disparate groups.

The conventional way to measure the association between a biomolecular pair is based on correlation (ρ_{ij}). A connection will be built when $|\rho_{ij}^{(1)} - \rho_{ij}^{(2)}|$ is statistically significant away from zero, where the superscript indicates the group index. As stated in Chapter 2, a drawback for using correlation is that correlation confounds direct and indirect associations [8]. For example, a strong correlation between x_1 and x_2 as well as x_2 and x_3 (direct associations) is very likely to introduce a relatively weak but still significantly strong correlation between x_1 and x_3 (indirect association).When the number of biomolecules is large, correlation tends to generate over-complicated networks, impacting the selection of reliable biomarker candidates in the subsequent analysis. Considering this, partial correlation that can distinguish direct and indirect associations will be helpful in generating a sparse differential network that are beneficial for both network visualization and reliable biomarker candidate selection.

Given a differential network, an straightforward way to select biomarker candidate is based on the node degree (i.e., the number of connections for each node) as what we have done in dwgLASSO [80]. The assumption is that biomolecules that have a strongly altered connectivity between biologically disparate groups might play an important role in the disease under study [13]. While the underlying assumption seems reasonable, this simple method does not consider the changes on expression levels of individual biomolecules between distinct biological groups. In fact, differential expression and differential network analyses investigate omic data from two separate but complementary perspectives: the former focuses on the change of single biomolecule in its mean expression level while the latter concentrates on the change in pairwise association for a biomolecular pair. Therefore, an approach that can integrate differential expression and differential network analyses is likely to discover more reliable biomarkers by considering the difference between distinct biological groups on both single biomolecule and biomolecular pair levels.

In this Chapter, we propose a novel approach, INDEED (INtegrated DiffErential Expression and Differential network analysis), to integrate differential expression and differential network analyses for biomarker discovery as shown in Figure 5.1. Given a single omic dataset, differential expression analysis is first performed to obtain p-value, which indicates the change of single biomolecule between distinct biological groups. Then, a differential network is built based on partial correlation, which can distinguish between direct and indirect associations when evaluating the change of pairwise association on a biomolecular pair between distinct biological groups. Activity scores are computed based on p-values and the topology of the differential network. Finally, biomolecules are prioritized by their activity scores for biomarker candidate selection. We show the application of INDEED through proteomic and glycomic data we previously acquired in our liver cancer biomarker discovery studies [81, 82]. We also apply INDEED on transcriptomic data we downloaded from online repository for breast cancer study [69, 83].

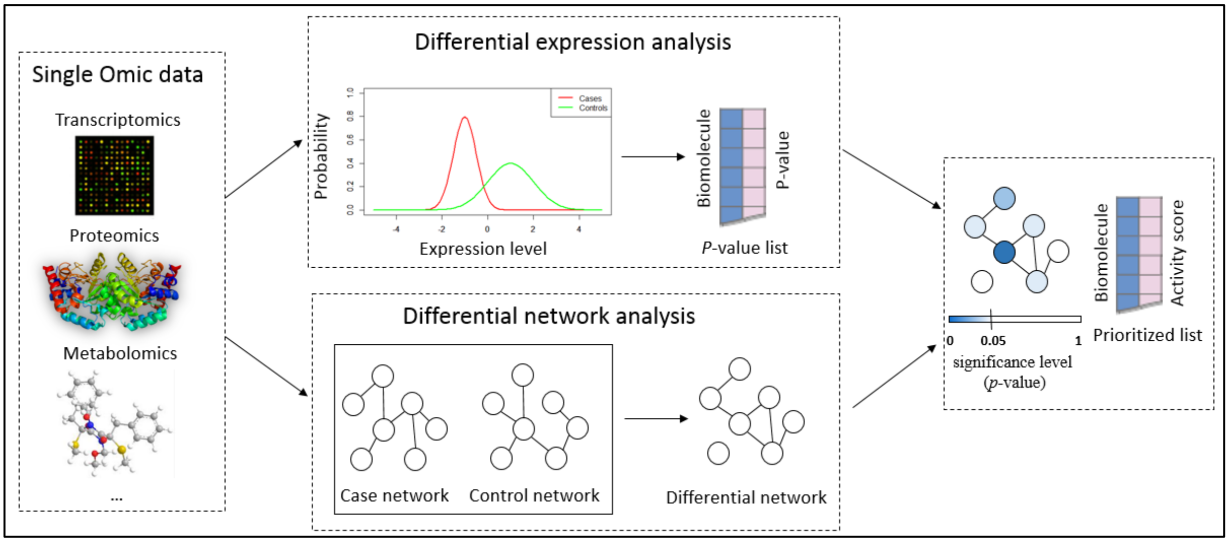


Figure 5.1 An overview of INDEED. The input is data matrix of one omic type (e.g., transcriptomics, proteomics, metabolomics) and the output is a prioritized list based on the activity score defined within INDEED.

Additionally, more insights can be gained through investigating various biological networks acting at different levels of human biological system (e.g., genes, proteins, metabolites, etc.). With multi-omic data for the same set of samples available, a method that can simultaneously integrate differential expression and differential network analyses, and take advantage of the information provided from multiple cellular components is promising to increase our understanding of cancer and to identify more powerful biomarkers. We extend INDEED as INDEED-M for integrative analysis of multi-omic data, and show its application using proteomic, metabolomics and glycomic data from the same set of samples with a focus on metabolite cancer biomarker discovery.

The rest of the Chapter is organized as follows. Section 5.2 introduces INDEED. Section 5.3 presents the performance of INDEED on real proteomic, glycomic, and transcriptomic data. In Section 5.4, we extend INDEED as INDEED-M for multi-omic data integration. Section 5.5 shows the application of INDEED-M using proteomic, metabolomic and glycomic data on the same set of samples. Finally, Section 5.6 summarizes our work.

5.2 Integrated differential expression and differential network analysis using INDEED

Figure 5.2 shows the framework of INDEED. It includes four steps: 1), performing differential expression analysis (e.g., student's *t*-test) to obtain p-value for each biomolecule; 2), building a differential network by evaluating the changes in partial correlation for each biomolecular pair between distinct biological groups; 3), computing the activity score for each biomolecule based on p-values from differential expression analysis and the topology of differential network; 4), prioritizing the biomolecules with the activity score.

Specifically, in step 1, differential expression analysis is typically performed through student's *t*-test, ANOVA, logistic regression or LASSO based method. Its aim is to detect the change in the expression level (i.e., p-value) of a single biomolecule between distinct biological groups.

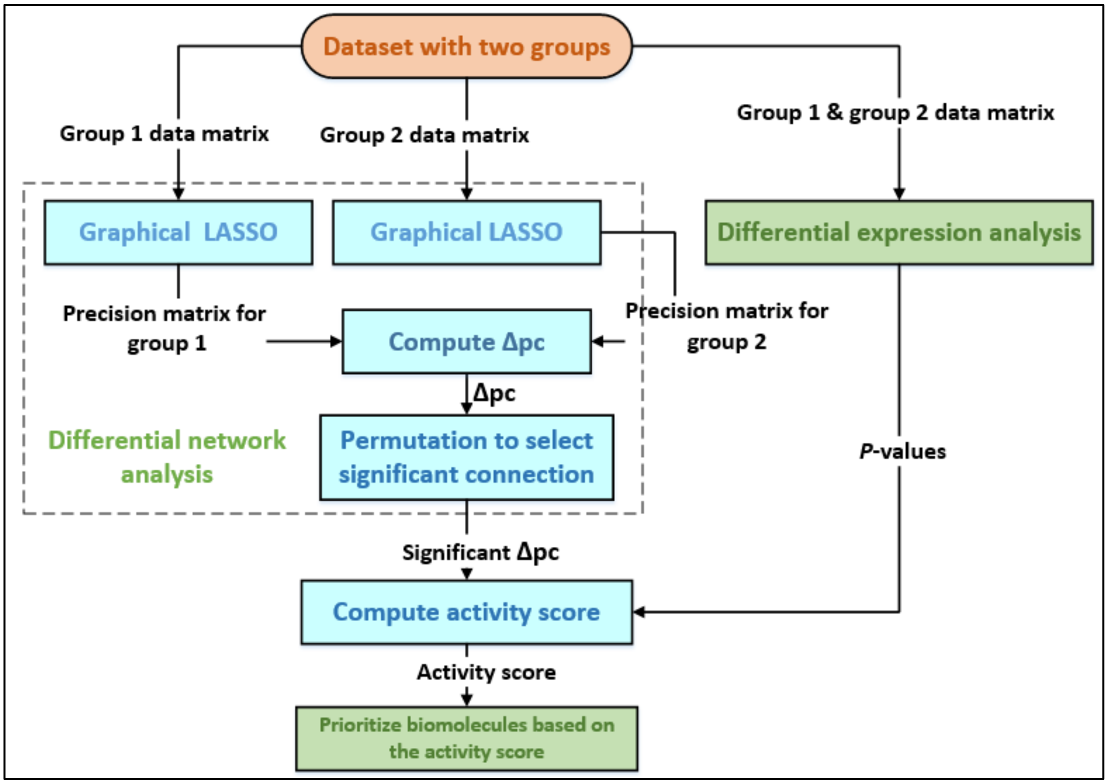

Dataset with two groups
Group 1 data matrix
Group 2 data matrix
Group 1 & group 2 data matrix
Graphical LASSO
Graphical LASSO
Differential expression analysis
Precision matrix for group 1
Compute Δpc
Precision matrix for group 2
Δpc
Differential network analysis
Permutation to select significant connection
P-values
Significant Δpc
Compute activity score
Activity score
Prioritize biomolecules based on the activity score

Figure 5.2 The framework of INDEED. In differential network analysis, the network is built based on partial correlation (pc).

In step 2, we build a differential network. Unlike the conventional way of using correlation to measure the pairwise association, we can obtain a sparse differential network by using partial correlation. This is due to the fact that conventional correlation confounds direct and indirect associations, while partial correlation can remove the effect of other biomolecules when evaluating a biomolecular pair [8, 43]. While correlation can be computed from covariance matrix, partial correlation can be computed from inverse covariance matrix (i.e., precision matrix $\boldsymbol{\Theta}$) as shown in Equation 5.1 [8].

$$pc_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \tag{5.1}$$

where pc_{ij} represents the partial correlation between x_i and x_j, and $\theta_{ij} \in \boldsymbol{\Theta}$.

Due to the ‘large *p* small *n*’ problem in omic data, the precision matrix $\mathbf{\Theta}$ is non-trivial to compute since the covariance matrix is singular. Graphical LASSO algorithm is widely used to efficiently estimate $\mathbf{\Theta}$ by solving the following optimization problem shown in Equation 5.2 [34, 84].

$$\arg\min_{\mathbf{\Theta}\succ\mathbf{0}} -\log\det\mathbf{\Theta} + \mathrm{tr}(\mathbf{S\Theta}) + \lambda\|\mathbf{\Theta}\|_1 \tag{5.2}$$

where $\mathbf{\Theta} \succ \mathbf{0}$ is the constraint that $\mathbf{\Theta}$ has to be positive definite, $\mathbf{S}$ is the sample covariance matrix, tr denotes trace, the sum of the diagonal elements in a matrix, $\|\mathbf{\Theta}\|_1$ represents the ℓ_1 norm of $\mathbf{\Theta}$, the sum of the absolute values of all the elements in $\mathbf{\Theta}$, and λ is the tuning parameter controlling the sparsity of $\mathbf{\Theta}$.

We perform graphical LASSO on distinct biological groups to obtain group-specific precision matrices (i.e., $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$). The sparsity parameters λ_1 and λ_2 in graphical LASSO as shown in Equations 5.3.1 and 5.3.2 are tuned by cross validation using one standard error rule. By applying one standard error rule, we can achieve the simplest (most regularized) model whose error is within one standard deviation of the minimal error. Based on our experience, other techniques such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and stability selection [49], are either prone to data under-fitting, leading to very large λ (e.g., AIC, BIC) or computationally very intensive (e.g., stability selection).

To be more specific, in step 2, we first perform graphical LASSO to build group-specific network. The sparsity parameters λ_1 and λ_2 in graphical LASSO as shown in Equation 5.1 are tuned by cross validation using one standard error rule.

$$\arg\min_{\mathbf{\Theta}_1\succ\mathbf{0}} -\log\det\mathbf{\Theta}_1 + \mathrm{tr}(\mathbf{S}_1\mathbf{\Theta}_1) + \lambda_1\|\mathbf{\Theta}_1\|_1 \tag{5.3.1}$$

$$\arg\min_{\mathbf{\Theta}_2 \succ \mathbf{0}} -\log\det\mathbf{\Theta}_2 + \mathrm{tr}(\mathbf{S}_2\mathbf{\Theta}_2) + \lambda_2\|\mathbf{\Theta}_2\|_1 \tag{5.3.2}$$

From the group-specific precision matrices $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$, we compute the partial correlation for each biomolecular pair in distinct biological groups $pc_{ij}^{(1)}$ and $pc_{ij}^{(2)}$ as shown in Equation 5.4 [8].

$$pc_{ij}^{(1)} = -\frac{\theta_{ij}^{(1)}}{\sqrt{\theta_{ii}^{(1)}\theta_{jj}^{(1)}}}, pc_{ij}^{(2)} = -\frac{\theta_{ij}^{(2)}}{\sqrt{\theta_{ii}^{(2)}\theta_{jj}^{(2)}}} \tag{5.4}$$

The change for each biomolecular pair in partial correlations between distinct biological groups is calculated as shown in Equation 5.5.

$$\Delta pc_{ij} = pc_{ij}^{(1)} - pc_{ij}^{(2)} \tag{5.5}$$

To evaluate the statistical significance of $\Delta pc_{ij} \neq 0$, we conduct a permutation test by randomly permuting the sample labels in distinct biological groups for each biomolecule, applying graphical LASSO under the same sparsity parameters previously used $\tilde{\lambda}_1 = \lambda_1$, $\tilde{\lambda}_2 = \lambda_2$, and finally computing $\widetilde{pc}_{ij}^{(1)}$, $\widetilde{pc}_{ij}^{(2)}$, and $\Delta\widetilde{pc}_{ij}$. This procedure is repeated 1000 times to obtain an empirical distribution of $\Delta\widetilde{pc}_{ij}$. $\Delta pc_{ij} \neq 0$ is considered statistically significant if Δpc_{ij} falls into the 2.5% tails on either end of the empirical distribution curve for $\Delta\widetilde{pc}_{ij}$. To build a differential network, we assign a connection between x_i and x_j when $\Delta pc_{ij} \neq 0$ is statistically significant.

In step 3, p-value (p_k) for each biomolecule is converted into z-score (z_k) as shown in Equation 5.6. An activity score (s_k) is defined as the summation of z_k and the z-scores for all its neighbors in the differential network, as shown in Equation 5.7. A higher activity score indicates that the corresponding biomolecule has more neighbors connected in the differential network and their p-values are more statistically significant.

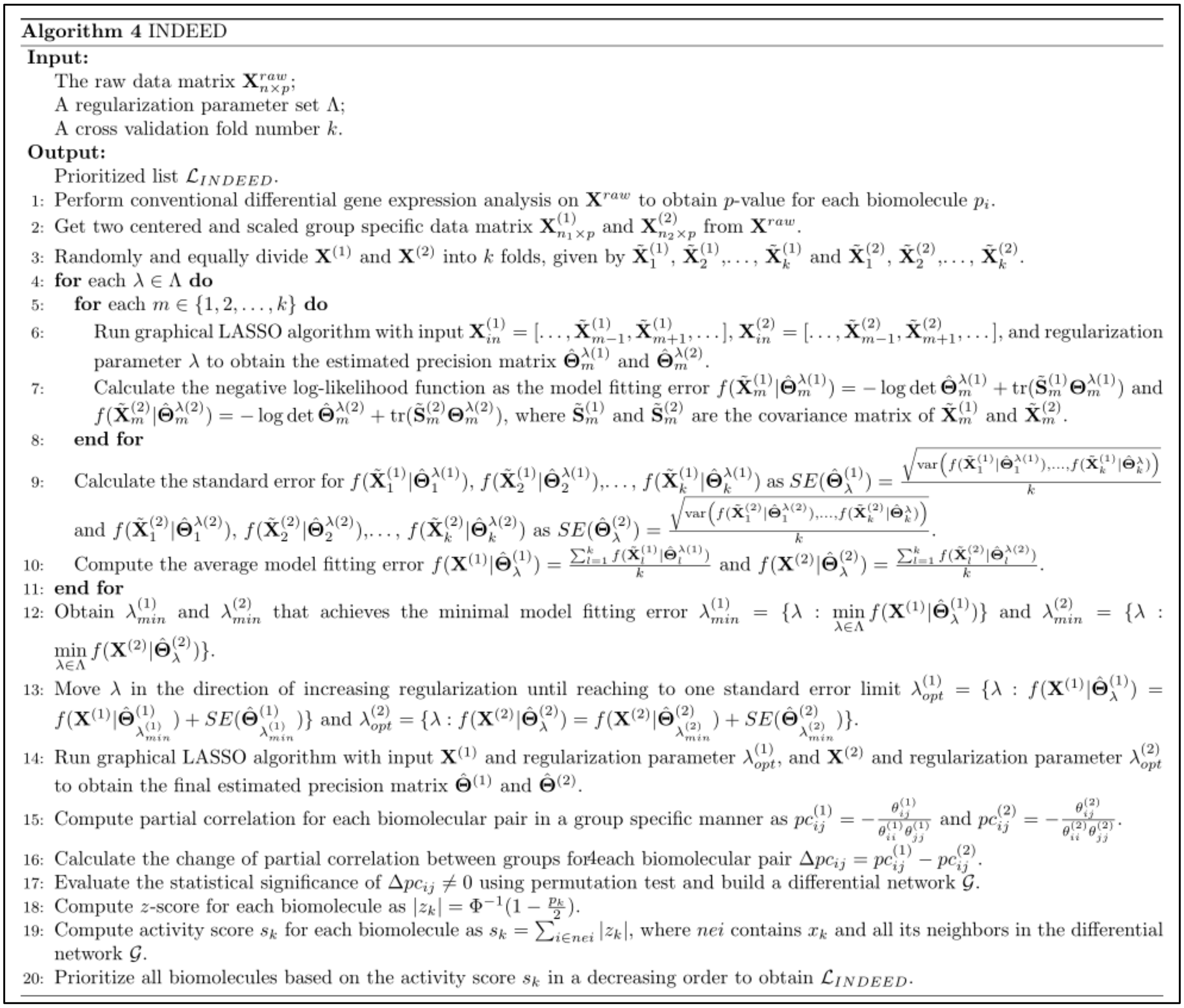
$$|z_k| = \phi^{-1}(1 - \tfrac{p_k}{2}) \tag{5.6}$$

where ϕ^{-1} is the inverse cumulative distribution function of the standard Gaussian distribution.

$$s_k = \sum_{i \in nei} |z_i| \tag{5.7}$$

where nei indicates x_k and its neighbors in the differential network.

Finally, in step 4, biomolecules are prioritized based on the activity score s_k and the top ranking biomolecules are selected as biomarker candidates. Detailed algorithm for INDEED is shown below:



```
Algorithm 4 INDEED
Input:
  The raw data matrix $\mathbf{X}^{raw}_{n\times p}$;
  A regularization parameter set $\Lambda$;
  A cross validation fold number $k$.
Output:
  Prioritized list $\mathcal{L}_{INDEED}$.
Perform conventional differential gene expression analysis on $\mathbf{X}^{raw}$ to obtain $p$-value for each biomolecule $p_i$.
Get two centered and scaled group specific data matrix $\mathbf{X}^{(1)}_{n_1\times p}$ and $\mathbf{X}^{(2)}_{n_2\times p}$ from $\mathbf{X}^{raw}$.
Randomly and equally divide $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ into $k$ folds, given by $\tilde{\mathbf{X}}^{(1)}_1$, $\tilde{\mathbf{X}}^{(1)}_2$,$\ldots$, $\tilde{\mathbf{X}}^{(1)}_k$ and $\tilde{\mathbf{X}}^{(2)}_1$, $\tilde{\mathbf{X}}^{(2)}_2$,$\ldots$, $\tilde{\mathbf{X}}^{(2)}_k$.
for each $\lambda \in \Lambda$ do
  for each $m \in \{1, 2, \ldots, k\}$ do
    Run graphical LASSO algorithm with input $\mathbf{X}^{(1)}_{in} = [\ldots, \tilde{\mathbf{X}}^{(1)}_{m-1}, \tilde{\mathbf{X}}^{(1)}_{m+1}, \ldots]$, $\mathbf{X}^{(2)}_{in} = [\ldots, \tilde{\mathbf{X}}^{(2)}_{m-1}, \tilde{\mathbf{X}}^{(2)}_{m+1}, \ldots]$, and regularization parameter $\lambda$ to obtain the estimated precision matrix $\hat{\mathbf{\Theta}}^{\lambda(1)}_m$ and $\hat{\mathbf{\Theta}}^{\lambda(2)}_m$.
    Calculate the negative log-likelihood function as the model fitting error $f(\tilde{\mathbf{X}}^{(1)}_m|\hat{\mathbf{\Theta}}^{\lambda(1)}_m) = -\log\det\hat{\mathbf{\Theta}}^{\lambda(1)}_m + \mathrm{tr}(\tilde{\mathbf{S}}^{(1)}_m\mathbf{\Theta}^{\lambda(1)}_m)$ and $f(\tilde{\mathbf{X}}^{(2)}_m|\hat{\mathbf{\Theta}}^{\lambda(2)}_m) = -\log\det\hat{\mathbf{\Theta}}^{\lambda(2)}_m + \mathrm{tr}(\tilde{\mathbf{S}}^{(2)}_m\mathbf{\Theta}^{\lambda(2)}_m)$, where $\tilde{\mathbf{S}}^{(1)}_m$ and $\tilde{\mathbf{S}}^{(2)}_m$ are the covariance matrix of $\tilde{\mathbf{X}}^{(1)}_m$ and $\tilde{\mathbf{X}}^{(2)}_m$.
  end for
  Calculate the standard error for $f(\tilde{\mathbf{X}}^{(1)}_1|\hat{\mathbf{\Theta}}^{\lambda(1)}_1)$, $f(\tilde{\mathbf{X}}^{(1)}_2|\hat{\mathbf{\Theta}}^{\lambda(1)}_2)$,$\ldots$, $f(\tilde{\mathbf{X}}^{(1)}_k|\hat{\mathbf{\Theta}}^{\lambda(1)}_k)$ as $SE(\hat{\mathbf{\Theta}}^{(1)}_\lambda) = \frac{\sqrt{\mathrm{var}\left(f(\tilde{\mathbf{X}}^{(1)}_1|\hat{\mathbf{\Theta}}^{\lambda(1)}_1),\ldots,f(\tilde{\mathbf{X}}^{(1)}_k|\hat{\mathbf{\Theta}}^{\lambda}_k)\right)}}{k}$ and $f(\tilde{\mathbf{X}}^{(2)}_1|\hat{\mathbf{\Theta}}^{\lambda(2)}_1)$, $f(\tilde{\mathbf{X}}^{(2)}_2|\hat{\mathbf{\Theta}}^{\lambda(2)}_2)$,$\ldots$, $f(\tilde{\mathbf{X}}^{(2)}_k|\hat{\mathbf{\Theta}}^{\lambda(2)}_k)$ as $SE(\hat{\mathbf{\Theta}}^{(2)}_\lambda) = \frac{\sqrt{\mathrm{var}\left(f(\tilde{\mathbf{X}}^{(2)}_1|\hat{\mathbf{\Theta}}^{\lambda(2)}_1),\ldots,f(\tilde{\mathbf{X}}^{(2)}_k|\hat{\mathbf{\Theta}}^{\lambda}_k)\right)}}{k}$.
  Compute the average model fitting error $f(\mathbf{X}^{(1)}|\hat{\mathbf{\Theta}}^{(1)}_\lambda) = \frac{\sum_{l=1}^{k} f(\tilde{\mathbf{X}}^{(1)}_l|\hat{\mathbf{\Theta}}^{\lambda(1)}_l)}{k}$ and $f(\mathbf{X}^{(2)}|\hat{\mathbf{\Theta}}^{(2)}_\lambda) = \frac{\sum_{l=1}^{k} f(\tilde{\mathbf{X}}^{(2)}_l|\hat{\mathbf{\Theta}}^{\lambda(2)}_l)}{k}$.
end for
Obtain $\lambda^{(1)}_{min}$ and $\lambda^{(2)}_{min}$ that achieves the minimal model fitting error $\lambda^{(1)}_{min} = \{\lambda : \min_{\lambda\in\Lambda} f(\mathbf{X}^{(1)}|\hat{\mathbf{\Theta}}^{(1)}_\lambda)\}$ and $\lambda^{(2)}_{min} = \{\lambda : \min_{\lambda\in\Lambda} f(\mathbf{X}^{(2)}|\hat{\mathbf{\Theta}}^{(2)}_\lambda)\}$.
Move $\lambda$ in the direction of increasing regularization until reaching to one standard error limit $\lambda^{(1)}_{opt} = \{\lambda : f(\mathbf{X}^{(1)}|\hat{\mathbf{\Theta}}^{(1)}_\lambda) = f(\mathbf{X}^{(1)}|\hat{\mathbf{\Theta}}^{(1)}_{\lambda^{(1)}_{min}}) + SE(\hat{\mathbf{\Theta}}^{(1)}_{\lambda^{(1)}_{min}})\}$ and $\lambda^{(2)}_{opt} = \{\lambda : f(\mathbf{X}^{(2)}|\hat{\mathbf{\Theta}}^{(2)}_\lambda) = f(\mathbf{X}^{(2)}|\hat{\mathbf{\Theta}}^{(2)}_{\lambda^{(2)}_{min}}) + SE(\hat{\mathbf{\Theta}}^{(2)}_{\lambda^{(2)}_{min}})\}$.
Run graphical LASSO algorithm with input $\mathbf{X}^{(1)}$ and regularization parameter $\lambda^{(1)}_{opt}$, and $\mathbf{X}^{(2)}$ and regularization parameter $\lambda^{(2)}_{opt}$ to obtain the final estimated precision matrix $\hat{\mathbf{\Theta}}^{(1)}$ and $\hat{\mathbf{\Theta}}^{(2)}$.
Compute partial correlation for each biomolecular pair in a group specific manner as $pc^{(1)}_{ij} = -\frac{\theta^{(1)}_{ij}}{\theta^{(1)}_{ii}\theta^{(1)}_{jj}}$ and $pc^{(2)}_{ij} = -\frac{\theta^{(2)}_{ij}}{\theta^{(2)}_{ii}\theta^{(2)}_{jj}}$.
Calculate the change of partial correlation between groups for4each biomolecular pair $\Delta pc_{ij} = pc^{(1)}_{ij} - pc^{(2)}_{ij}$.
Evaluate the statistical significance of $\Delta pc_{ij} \neq 0$ using permutation test and build a differential network $\mathcal{G}$.
Compute $z$-score for each biomolecule as $|z_k| = \Phi^{-1}(1 - \frac{p_k}{2})$.
Compute activity score $s_k$ for each biomolecule as $s_k = \sum_{i\in nei}|z_k|$, where $nei$ contains $x_k$ and all its neighbors in the differential network $\mathcal{G}$.
Prioritize all biomolecules based on the activity score $s_k$ in a decreasing order to obtain $\mathcal{L}_{INDEED}$.
```

5.3 Evaluation of INDEED on various omic data

5.3.1 Evaluation of INDEED using proteomic data

The proteomic datasets were acquired by analysis of proteins in sera from hepatocellular carcinoma (HCC) cases and liver cirrhotic controls [81]. Briefly, adult patients were recruited from MedStar Georgetown University Hospital (GU cohort) in Washington, DC, USA and the Tanta University Hospital (TU cohort) in Tanta, Egypt. The GU cohort is comprised of 116 subjects (57 HCC cases and 59 liver cirrhotic controls) and the TU cohort consists of 89 subjects

(40 HCC cases and 49 liver cirrhotic controls). We used liquid chromatography coupled with mass spectrometry (LC-MS) for both untargeted and targeted analyses of sera from subjects in the GU and TU cohorts. Proteins that are statistically significant between the two groups were selected from the untargeted proteomic data. A total of 101 proteins were then evaluated in sera from the GU and TU cohorts through targeted quantitation using multiple reaction monitoring (MRM). More details on experiment design and statistical analysis can be found in [81].

Our goal is to obtain a prioritized list of proteins using INDEED in one cohort, select the top ranking proteins to build a disease classifier and evaluate the performance of these proteins and the classifier on the other cohort with independent subjects. GU cohort was used as the training set for the selection of proteins and the built of the classifier, since it has more subjects and almost the same number of HCC cases and liver cirrhotic controls. In contrast, TU cohort was used as the testing set.

We performed student's t-test on the GU cohort to investigate the changes on the expression level of individual proteins between HCC cases and liver cirrhotic controls. For each protein, we obtained a p-value (p_k) from student's t-test. The group-specific matrix (i.e., HCC cases or liver cirrhotic controls) from GU cohort was then used as the input for graphical LASSO algorithm to obtain the group-specific precision matrices ($\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ for HCC and cirrhotic groups, respectively). In graphical LASSO, we performed 5-fold cross validation and chose the optimal tuning parameter λ in Equation 5.3.1 and 5.3.2 by one standard error rule as shown in Figure 5.3.

From $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$, we computed the partial correlation for each biomolecular pair in HCC and cirrhotic groups $pc_{ij}^{(1)}$ and $pc_{ij}^{(2)}$ (Equation 5.4) and the change for pairwise partial correlation between the two groups Δpc_{ij} (Equation 5.5). To evaluate the statistical significance of $\Delta pc_{ij} \neq 0$, we conducted permutation test as explained in Section 5.2.1. To build a differential network, we assigned a connection between x_i and x_j, when $\Delta pc_{ij} \neq 0$ is statistically significant.

We mapped the p-values (p_k) for each protein onto the differential network as shown in Figure 5.4, computed the activity score (s_k) for each protein, as defined in Equations 5.6 and 5.7, and prioritized the 101 proteins according to their activity scores in a decreasing order.

To evaluate the performance of INDEED, we also prioritized the 101 proteins according to differential expression analysis (i.e., the p-values from student's t-test) and differential network analysis. In differential network analysis, we used the differential network in Figure 5.4 and prioritized the proteins according to the node degree of each protein (i.e., how many neighbors one node is connected to). The top ranking proteins from the three prioritized lists were used to train three logistic regression classifiers and tested their performances on the independent testing dataset.

Figure 5.3 shows our choice of $\lambda_1 = 0.106$ (HCC group) and $\lambda_2 = 0.125$ (cirrhotic group) in performing graphical LASSO to obtain group-specific precision matrices ($\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ for HCC and cirrhotic groups, respectively).

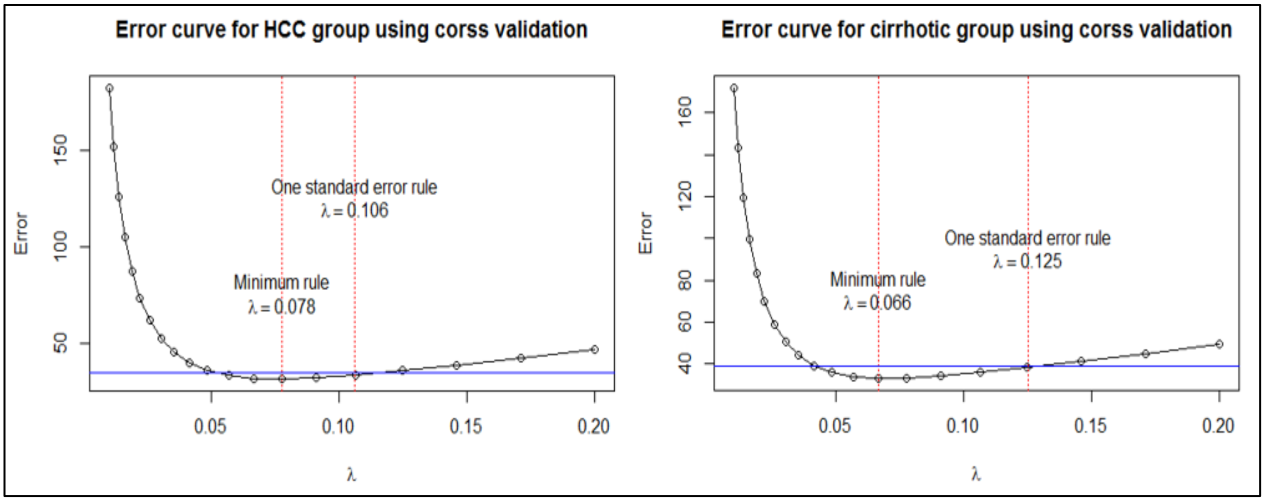


Figure 5.3 Error curves to choose optimal tuning parameter λ using 5-fold cross validation by one standard error rule for HCC and cirrhotic groups on proteomic data. The blue line indicates the one standard error for the minimum λ in the direction of increasing regularization.

The differential network built based on partial correlation is shown in Figure 5.4. Table S-1 in supplementary material lists all the 101 proteins together with their adjusted *p*-values, activity scores and node degrees [85]. Proteins are named after their corresponding gene symbols.

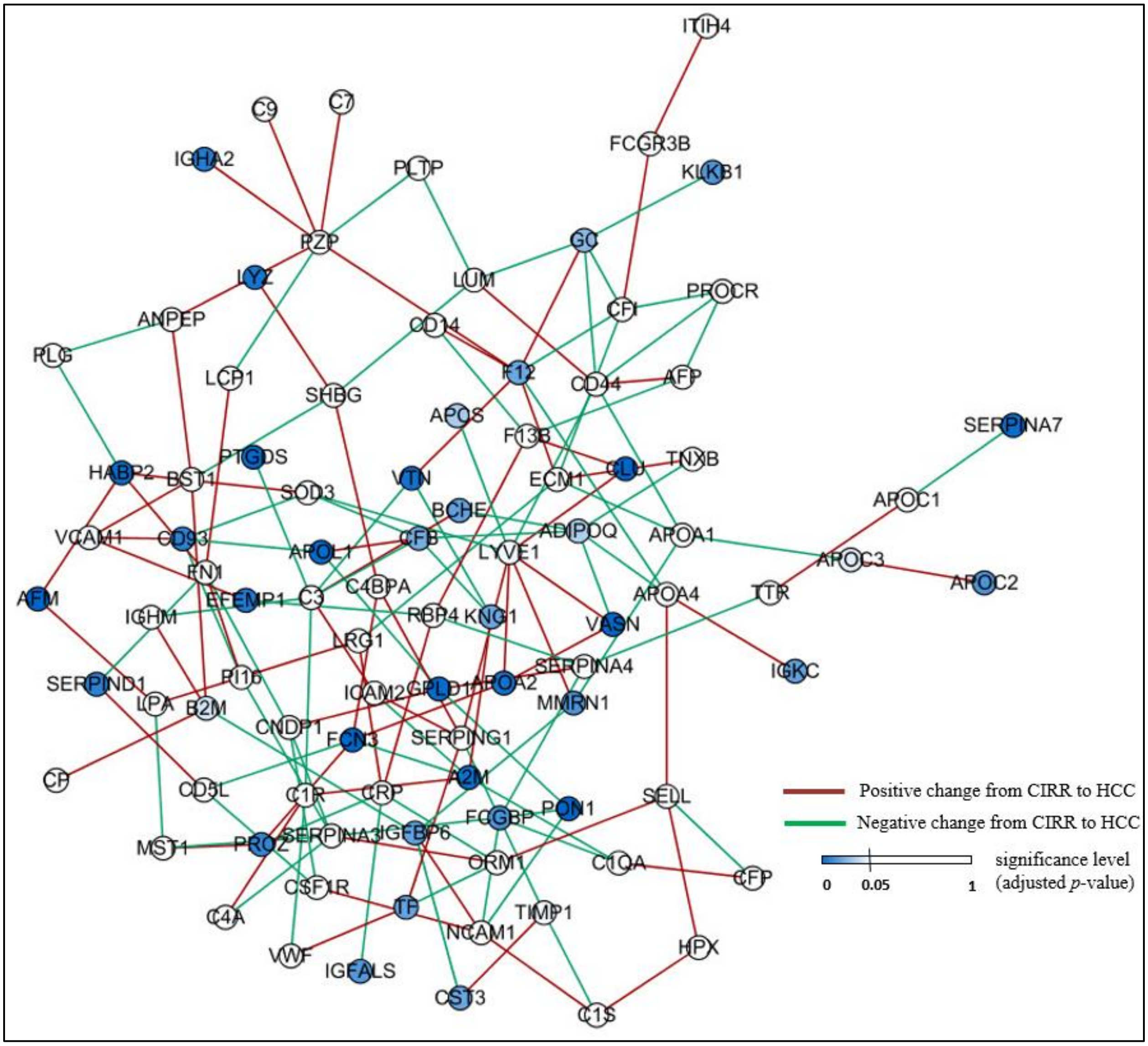


Figure 5.4 Differential network from proteomic data. Node color indicates the significance level of the individual protein between the HCC and cirrhotic groups. Orange edge represents a significantly positive change on partial correlation (pc) of a protein pair from cirrhotic to HCC groups while green one indicates a significantly negative change.

We performed differential expression analysis, differential network analysis, and INDEED on GU cohort initially. Using student's t-test, 45 proteins with adjusted p-values less than 0.05 were selected in differential expression analysis. The inflation of Type I error is controlled by the false discovery rate (FDR) using the Benjamini-Hochberg procedure. To make a fair comparison, we also selected the top 45 proteins based on differential network analysis (i.e., node degrees) and INDEED (i.e., activity scores). We conducted student's t-test on the TU cohort to select a total of 39 proteins whose adjusted p-values were less than 0.05. We compared the overlap of the 45 proteins selected based on differential expression analysis, differential network analysis and INDEED on GU cohort, with the 39 proteins selected by student's t-test on the TU cohort. The result is shown in Table 5.1, where the number of overlapping proteins are 21, 17, and 25 for differential expression analysis, differential network analysis and INDEED, respectively. Here the 39 proteins selected by student's t-test on the TU cohort are used to approximate the ground truth to evaluate the reproducibility of the protein biomarker candidates selected based on differential expression analysis, differential network analysis and INDEED from GU cohort. As expected, INDEED can select biomarker candidates that are more reproducible across GU and TU cohorts.

Table 5.1 The top ranking 45 proteins prioritized by differential expression (DE) analysis (adjusted p-value < 0.05), differential network (DN) analysis and INDEED on GU cohort. The reference is the top ranking proteins prioritized by DE analysis (adjusted p-value < 0.05) on TU cohort. The overlapping proteins between the three prioritized lists on GU cohort and the reference on TU cohort are in bold. All proteins are represented by their corresponding gene symbols.

GU cohort						TU cohort	
DE analysis *(Overlap:21)*		DN analysis *(Overlap:17)*		INDEED *(Overlap:25)*		DE analysis	
FCN3	IGFALS	**LYVE1**	**F13B**	**LYVE1**	FCGBP	PLG	C7
CLU	IGKC	F12	APOA4	**C3**	CD44	APCS	ADIPOQ
PON1	BCHE	**C3**	SOD3	FCN3	PON1	ICAM2	APOA2
AFM	IGFBP6	PZP	**SELL**	**CFB**	**F13B**	VCAM1	SERPIND1
VASN	COMP	SERPINA3	IGFBP6	SOD3	**VTN**	CFB	B2M
PTGDS	TF	CD44	**SERPINA4**	F12	**MMRN1**	TNXB	PTGDS
APOL1	F12	FCN3	**LUM**	**GPLD1**	FN1	SERPINA4	BST1
GPLD1	**CST3**	**CFB**	**B2M**	**ADIPOQ**	CD5L	C3	APOC2
SERPINA7	**KNG1**	A2M	**GPLD1**	**APOA2**	APOA4	KNG1	LUM
HABP2	IGJ	CRP	**BST1**	APOL1	**VCAM1**	CLU	VASN
A2M	**CFB**	FN1	HABP2	SERPING1	**PROZ**	VTN	APOC3
VTN	GC	GC	CNDP1	A2M	**ECM1**	F13B	CST3
APOA2	FGB	SERPING1	**CD93**	HABP2	APOA1	PROZ	C4BPA
LYZ	IGHA1	**CFI**	APOL1	**KNG1**	**CFI**	SELL	CD93
EFEMP1	**APCS**	NCAM1	CD5L	**CD93**	**CLU**	APOC1	CFI
IGHA2	**ADIPOQ**	**ADIPOQ**	TIMP1	NCAM1	**ICAM2**	RBP4	FGG
SERPIND1	CLEC3B	**ECM1**	**KNG1**	**VASN**	**B2M**	CSF1R	
CD93	**B2M**	FCGBP	**APOA2**	GC	**CST3**	GPLD1	
PROZ	**APOC3**	C1R	C1QA	IGFBP6	TIMP1	EFEMP1	
FCGBP	TIMP1	APOA1	LRG1	**RBP4**	**C4BPA**	MMRN1	
APOC2	**VCAM1**	**RBP4**	AFP	PZP	**APOC3**	ITIH4	
MMRN1	**LYVE1**	ORM1	TF	**SERPINA4**	**BST1**	ECM1	
KLKB1		SHBG		CRP		LYVE1	

Figure 5.5 shows a Venn diagram of the 21, 17, and 25 overlapping proteins selected by differential expression analysis, differential network analysis and INDEED from GU cohort. Two proteins, intercellular adhesion molecule 2 (ICAM2) and c4b-binding protein alpha chain (C4BPA) are unique to INDEED. We further investigated these two proteins by their relevance

to HCC studies from the past literatures. ICAM2 has been previously reported as a liver cirrhosis signature in plasma that can be used as a potential predictive biomarker for HCC among hepatitis B virus (HBV) carriers [86]. C4BPA has also been previously reported as one of the 14 protein biomarkers for HCC based on a study comparing HCC cases with healthy controls and the HBV group [87]. The literature survey has confirmed the prospective of using INDEED to select HCC related biomarker candidates that can be missed by differential expression and differential network analyses.

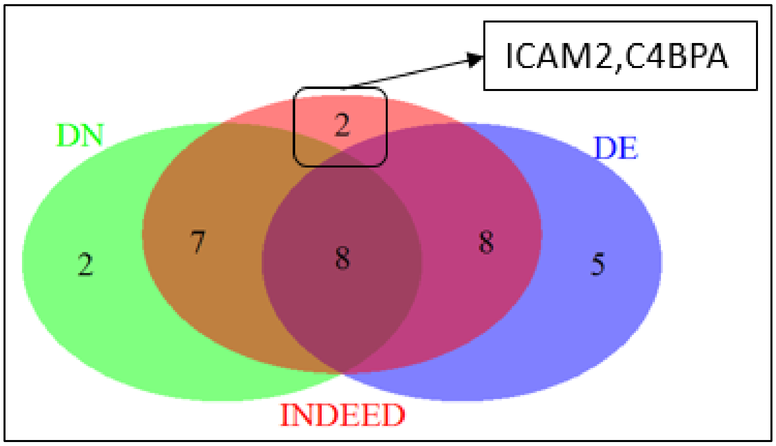


Figure 5.5 Venn diagram for the 21, 17 and 25 overlapping proteins from differential expression (DE) analysis, differential network (DN) analysis and INDEED on GU cohort in Table 5.1. Proteins ICAM2 and C4BPA are unique to INDEED.

To make more comprehensive comparisons among differential expression analysis, differential network analysis, and INDEED, we trained three logistic regression classifiers on GU cohort using the 45 proteins from differential expression analysis, differential network analysis and INDEED in Table 5.1, and tested the classifiers on the TU cohort. To overcome the potential over-fitting problem, we first performed a LASSO based logistic regression using R package, glmnet, to select the most relevant biomarker candidates among the 45 proteins in

Table 5.1 [88]. The sparsity parameter was tuned based on the leave-one-out cross validation procedure. This led to 10, 10, and 13 proteins for differential expression analysis, differential network analysis, and INDEED, respectively, as shown in Table 5.2. We then refitted the logistic regression classifiers using the above 10, 10, and 13 proteins and tested the classifiers on the TU cohort. The classification accuracy for the logistic regression classifiers on TU cohort are 0.64, 0.64, and 0.69 for differential expression analysis, differential network analysis and INDEED, respectively. We also plotted the ROC curves associated with differential expression analysis, differential network analysis and INDEED, as shown in Figure 5.6. The AUC for differential expression analysis, differential network analysis and INDEED are 0.68, 0.65 and 0.71, respectively.

Table 5.2 The 10, 10, 13 proteins selected by LASSO based logistic regression for differential expression (DE) analysis, differential network (DN) analysis and INDEED on GU cohort.

DE analysis (10)	DN analysis (10)	INDEED (13)
FCN3	F12	FCN3
IGHA2	FCN3	F12
CLEC3B	A2M	SERPING1
SERPINA7	SERPING1	A2M
CLU	FCGBP	VASN
PTGDS	IGFBP6	IGFBP6
AFM	GPLD1	FCGBP
LYZ	HABP2	PON1
VASN	TIMP1	APOA4
FCGBP	AFP	PROZ
		CLU
		B2M
		CST3

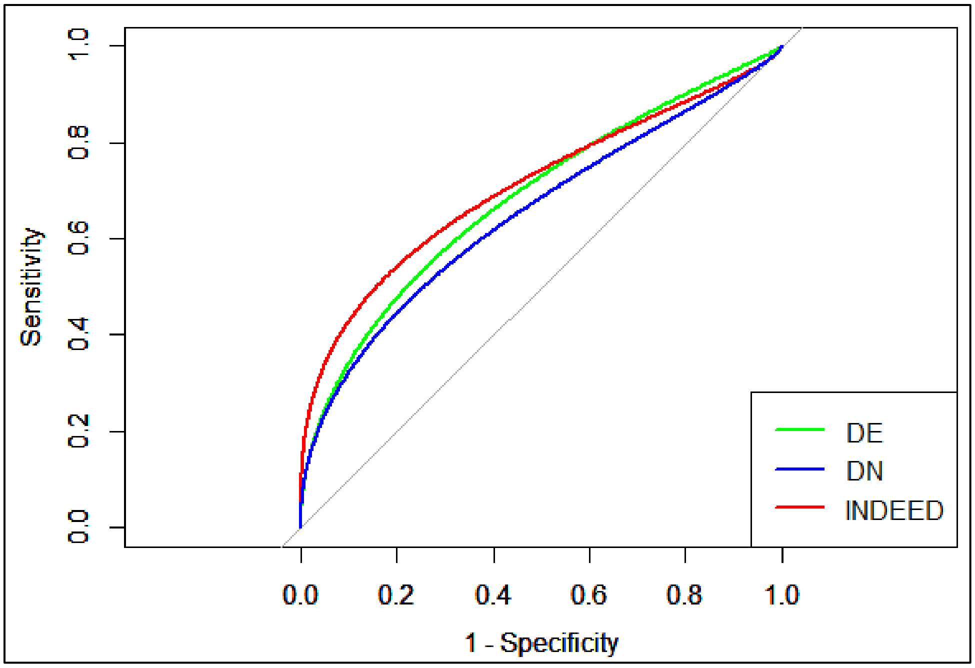


Figure 5.6 ROC curves associated with differential expression (DE) analysis, differential network (DN) analysis and INDEED when training a logistic regression classifier on GU cohort and testing it on TU cohort for proteomic data. The AUC are 0.68, 0.65 and 0.71 for DE analysis, DN analysis and INDEED, respectively.

5.3.2 Evaluation of INDEED using glycomic data

The glycomic datasets were acquired by analysis of glycans in sera from HCC cases and liver cirrhotic controls [82]. Similar to the proteomic datasets, adult patients were recruited from MedStar Georgetown University Hospital (GU cohort) in Washington, DC, USA and the Tanta University Hospital (TU cohort) in Tanta, Egypt. The GU cohort is comprised of 94 subjects (48 HCC cases and 46 patients with liver cirrhosis) and the TU cohort consists of 89 subjects (40 HCC cases and 49 liver cirrhotic controls). Both untargeted and targeted analyses were conducted by using LC-MS in the GU and TU cohorts. Glycans that are statistically significant between the HCC and cirrhotic groups were selected from the untargeted glycomic data. A total of 82 glycans were then evaluated in sera from the GU and TU cohorts through targeted quantitation using MRM. More details on experiment design and statistical analysis can be found in [82].

Similar to the approach we used for evaluating INDEED on proteomic data, we used GU cohort as the training set and TU cohort as the testing set. We performed ANOVA to investigate the changes on the expression level of individual glycans between HCC cases and liver cirrhotic controls. For each glycan, we obtained a p-value (p_k) from ANOVA. Then the differential network was built by performing graphical LASSO on HCC and cirrhotic groups separately, computing partial correlation for each glycan pair and evaluating the statistical significance of the change on the pairwise partial correlation between HCC and cirrhotic groups using permutation test. Once the differential network was built, we mapped p-values (p_k) onto the differential network and computed the activity score (s_k) for each glycan. At last, we prioritized the 82 glycans according to their activity scores in a decreasing order. To evaluate the performance of INDEED, we also prioritized the 82 glycans according to differential expression analysis (i.e., the p-values from ANOVA) and differential network analysis (i.e., node degrees). The top ranking glycans from the three prioritized lists were used to train three logistic regression classifiers. We tested the performances of the classifiers on the independent testing dataset.

Figure 5.7 shows our chose of $\lambda_1 = 0.066$ (HCC group) and $\lambda_2 = 0.057$ (cirrhotic group) in performing graphical LASSO to obtain group-specific precision matrices ($\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ for HCC and cirrhotic groups, respectively). Figure 5.8 shows the differential network built based on partial correlation. Table S-2 lists all 82 glycans together with their p-values, activity scores and node degrees [85].

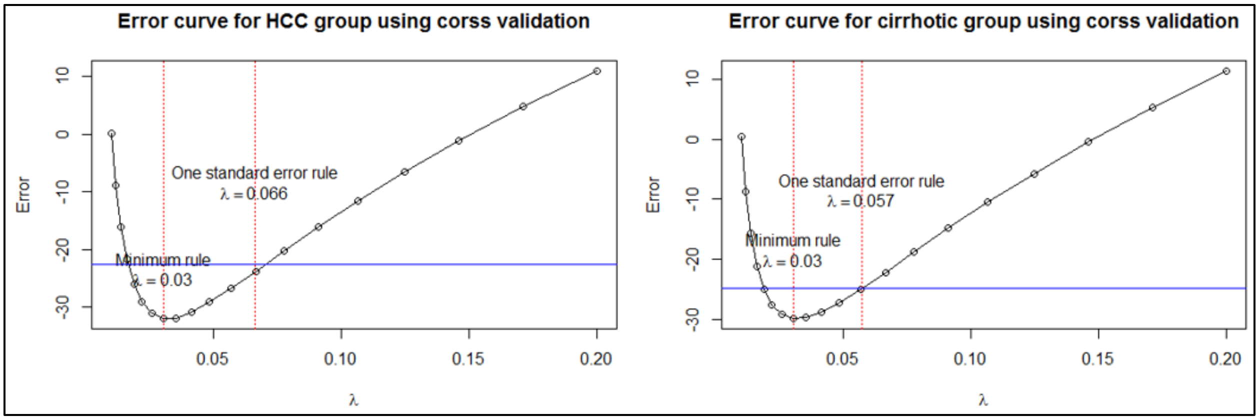
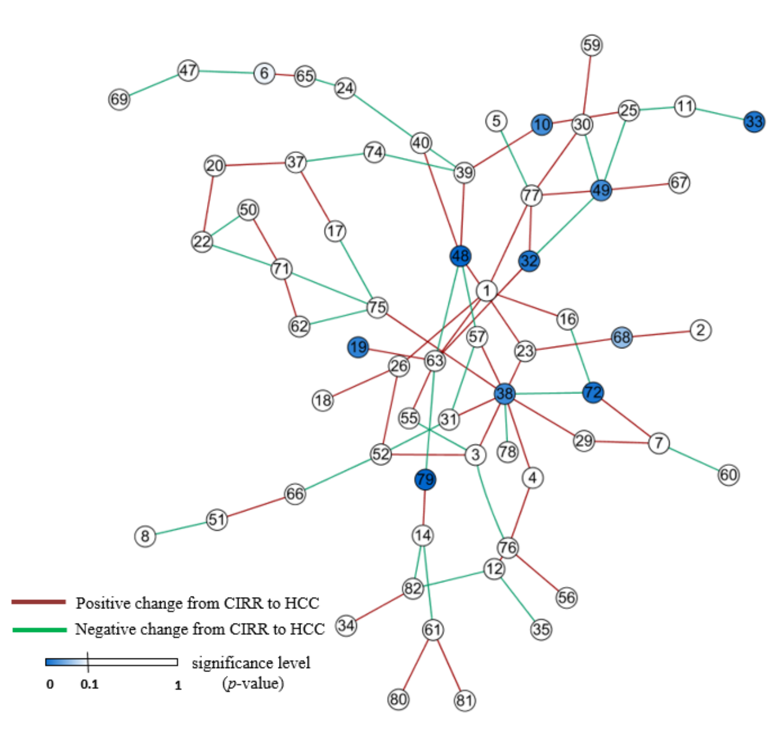


Figure 5.7 Error curves to choose optimal tuning parameter λ using 5-fold cross validation by one standard error rule for HCC and cirrhotic groups on glycomic data. The blue line indicates the one standard error for the minimum λ in the direction of increasing regularization.


Positive change from CIRR to HCC
Negative change from CIRR to HCC
significance level
(p-value)
0
0.1
1

ID	Glycan structure	Retention time (min)	ID	Glycan structure	Retention time (min)
1	53100	30.0	42	53321	30.5
2	25000	28.3	43	53321	37.5
3	43201	31.5	44	53321	30.5
4	43000	25.3	45	53321	37.5
5	43211	34.3	46	53321	30.5
6	53111	35.8	47	53302	34.3
7	53201	34.0	48	53411	34.3
8	26000	32.0	49	34101	32.3
9	53303	36.5	50	63423	43.0
10	34100	29.5	51	28000	39.0
11	34100	27.0	52	28000	41.5
12	43010	27.5	53	63404	40.5
13	63402	38.0	54	63404	36.3
14	43202	34.3	55	43101	30.0
15	43100	26.0	56	53312	37.8
16	43100	30.8	57	53312	42.0
17	53313	38.3	58	53110	33.3
18	53211	36.3	59	63414	42.3
19	53000	29.5	60	53200	31.5
20	53323	40.8	61	35011	34.3
21	43212	36.0	62	63424	44.0
22	43212	40.5	63	43111	31.5
23	33101	28.8	64	53303	36.5
24	33101	36.0	65	29000	42.3
25	43212	36.0	66	29000	45.0
26	34110	32.3	67	43201	31.5
27	27000	35.5	68	43201	31.5
28	63403	39.5	69	66012	39.5
29	63403	36.5	70	63402	38.0
30	53311	36.8	71	63434	47.0
31	53311	39.3	72	53101	33.5
32	43110	28.5	73	53210	34.3
33	43200	28.0	74	53313	38.3
34	43200	33.8	75	73514	38.3
35	53010	31.8	76	73514	41.0
36	53100	30.0	77	43211	34.3
37	63413	41.0	78	210000	47.8
38	53212	38.8	79	53111	35.8
39	53212	33.8	80	53220	36.8
40	53212	36.3	81	53201	34.0
41	53321	37.5	82	43202	34.3

Figure 5.8 Differential network from glycomic data. Node color indicates the significance level of the individual glycan between the HCC and cirrhotic groups. Orange edge represents a significantly positive change on partial correlation (pc) of a protein pair from cirrhotic to HCC groups while green one indicates a significantly negative change.

We performed differential expression analysis, differential network analysis, and INDEED on GU cohort. Using ANOVA, 11 glycans with p-values less than 0.1 were selected in differential expression analysis. To make a fair comparison, we selected the top 11 glycans based on differential network analysis (i.e., node degrees) and INDEED (i.e., activity scores) (Table 5.3). We then trained three logistic regression classifiers on GU cohort using the 11 glycans from differential expression analysis, differential network analysis and INDEED in Table 5.3, and tested the classifiers on the TU cohort. The same procedure as the proteomic data has been applied for the glycomic data. Briefly, we first performed a LASSO based logistic regression to select the most relevant biomarker candidates among the 11 glycans in Table 5.3. This led to 4, 2, and 5 glycans for differential expression analysis, differential network analysis, and INDEED, respectively, as shown in Table 5.4. We then refitted logistic regression classifiers using the above 4, 2, and 5 glycans and tested the classifiers on the TU cohort. The classification accuracy for the logistic regression classifiers on TU cohort are 0.58, 0.56, and 0.63 for differential expression analysis, differential network analysis and INDEED, respectively. We also plotted the ROC curves associated with differential expression analysis, differential network analysis, and INDEED, as shown in Figure 5.9. The AUC for differential expression analysis, differential network analysis and INDEED are 0.64, 0.59 and 0.67, respectively.

Table 5.3 The top 11 glycans prioritized based on differential expression (DE) analysis, differential network (DN) analysis and INDEED, respectively.

GU cohort					
DE analysis		DN analysis		INDEED	
Glycan structure	Retention time (min)	Glycan structure	Retention time (min)	Glycan structure	Retention time (min)
34100	29.5	53212	38.8	43111	31.5
43100	26	43111	31.5	53212	38.8
53000	29.5	53411	34.3	43211	34.3
43110	28.5	34101	32.3	34101	32.3
43200	28	43211	34.3	53212	33.8
53212	38.8	53100	30	53312	42
53411	34.3	43201	31.5	33101	28.8
34101	32.3	53212	33.8	53411	34.3
43201	31.5	28000	41.5	43202	34.3
53101	33.5	63434	47	53101	33.5
53111	35.8	73514	38.3	63434	47

Table 5.4 The 4, 2, 5 glycans selected by LASSO based logistic regression classifier for differential expression (DE) analysis, differential network (DN) analysis and INDEED on GU cohort. Glycans are characterized by the number of five monosaccharides: GlcNAc, mannose, galactose, fucose, and NeuNAc.

DE analysis (4)	DN analysis (2)	INDEED (5)
[43100]	[34101]	[53212]
[53000]	[53100]	[34101]
[53411]		[33101]
[53111]		[53411]
		[43202]

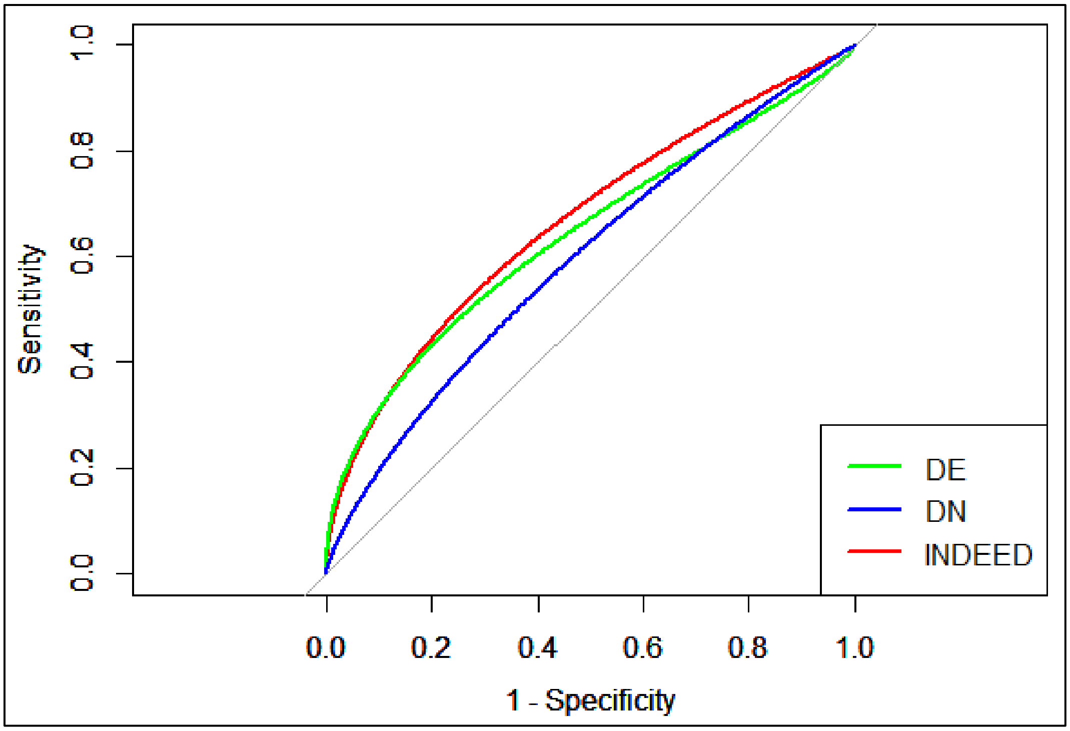


Figure 5.9 ROC curves associated with differential expression (DE) analysis, differential network (DN) analysis, and INDEED when training a logistic regression classifier on GU cohort and testing it on TU cohort for glycomic data. The AUC are 0.64, 0.59 and 0.67 for DE analysis, DN analysis and INDEED, respectively.

5.3.3 *Evaluation of INDEED using transcriptomic data*

The transcriptomic data consist of two microarray datasets previously acquired in breast cancer studies: van de Vijver *et al.*'s and Pawitan *et al.*'s datasets [69, 83]. The former includes 295 patients with their survival records, and was used for training. Pawitan *et al.*'s dataset contains 159 patients, together with their survival records, and was used for independent testing. Both datasets are available at PRECOG website (https://precog.stanford.edu/), an online repository for querying cancer gene expression and clinical data, and have been properly preprocessed for subsequent statistical analysis [89]. The raw data are also available at R package seventyGeneData and Gene Expression Omnibus (GSE1456), respectively [90].

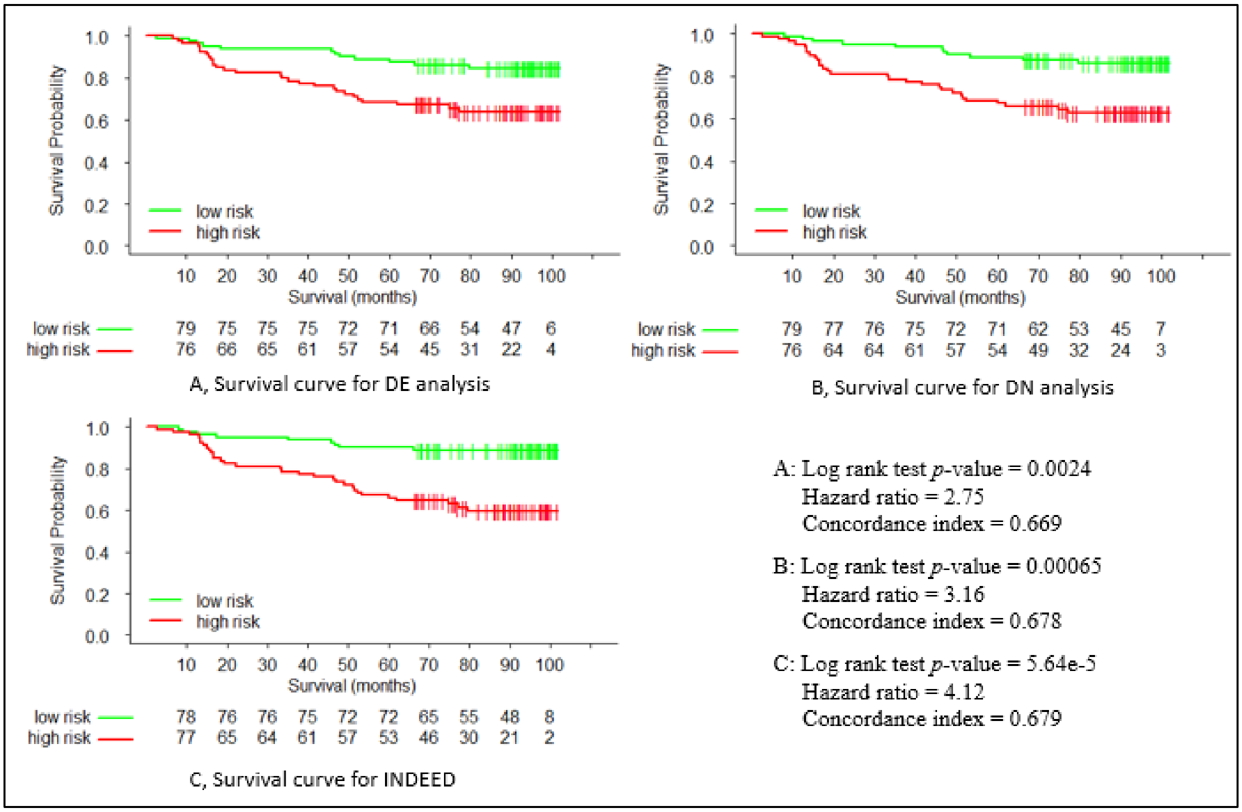
With proteomic and glycomic data in Sections 5.3.1 and 5.3.2, we evaluated INDEED by obtaining a prioritized list of proteins/glycans based on one dataset (i.e., GU cohort), selected top ranking ones to build a disease classifier, and tested the performance of the classifier on the independent dataset (i.e., TU cohort). For transcriptomic data, we evaluated INDEED by building a regression model for survival time prediction. We first conducted univariate analysis on van de Vijver *et al.*'s dataset to select a list of statistically significant genes based on their expression values and the survival time across patients using univariate Cox regression model. For each gene, we obtained a p-value (p_k) and selected statistically significant genes for subsequent analysis. In order to build a differential network, we excluded patients with less than 5-year follow-up time from the van de Vijver *et al.*'s dataset. Among the remaining patients, 91 with less than 5-year survival during the follow-up time were considered high risk group while the other 196 formed the low risk group. The differential network was built by performing graphical LASSO on high and low risk groups separately using the pre-selected genes, computing partial correlation for each gene pair, and evaluating the statistical significance of the change on the pairwise partial correlation between high and low risk groups using permutation test. Once the differential network was built, we mapped p-values (p_k) onto the differential network and computed the activity score (s_k) for each pre-selected gene. At last, we prioritized the pre-selected genes according to their activity scores in a decreasing order. To evaluate the performance of INDEED, we prioritized the pre-selected genes according to differential expression (i.e., the p-values from univariate Cox regression model) and differential network (i.e., node degrees) analyses. The top ranking genes from the three prioritized lists were used to train three multivariate Cox regression models and to test their performance on the independent testing dataset.

We performed univariate analysis on van de Vijver *et al.*'s dataset to select a list of statistically significant genes based on their expression value and the survival time across patients using univariate Cox regression model. This led to a total of 402 genes whose adjusted p-values were less than 0.05 after correcting for multiple testing based on FDR. Using cross-validation similar to Figures 5.3 and 5.7, we chose $\lambda_1 = 0.103$ (high risk group) and $\lambda_2 = 0.074$ (low risk group) in performing graphical LASSO to obtain group-specific precision matrices ($\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ for high and low risk groups, respectively). Table S-4 lists all 402 genes together with their p-values, activity scores, and node degrees [85].

We performed differential expression analysis, differential network analysis, and INDEED to prioritize the 402 genes based on their p-values, node degrees, and activity scores, respectively. From the three prioritized lists, the top 50 genes were selected to train three multivariate Cox regression models for survival time prediction. In training each multivariate Cox regression model, we used LASSO to select the most relevant biomarker candidates among the 50 genes. This led to 16, 23, and 22 genes selected by differential expression analysis, differential network analysis, and INDEED, respectively, as shown in Table 5.5. We then refitted the multivariate Cox regression models using the above 16, 23, and 22 genes and tested their performance on the independent Pawitan *et al.*'s dataset. Figure 5.10 presents survival curves associated with differential expression analysis, differential network analysis, and INDEED based on Kaplan-Meier survival analysis. As shown in the figure, INDEED yielded the best performance (log rank p-value=5.64e^{-5}, hazard ratio=4.12), compared to differential expression analysis (log rank p-value=0.0024, hazard ratio=2.75) and differential network analysis (log rank p-value=0.00065, hazard ratio=3.16).

Table 5.5 The 16, 23, 22 genes selected by LASSO based multivariate Cox regression models for differential expression (DE) analysis, differential network (DN) analysis and INDEED on van de Vijver *et al.*'s dataset.

DE analysis (16)	DN analysis (23)		INDEED (22)	
QSOX2	LRIG1	SPEF1	ZWINT	MED11
UBE2C	ZWINT	PLK2	CCNA2	ODF2
POLD1	MASTL	C20ORF24	SIK3	DSCR6
BIRC5	CSNK1D	TBC1D8	LZTFL1	NEIL1
PSMA7	CHMP1A	DSCR6	ABCB6	JMJD1C
SPC25	STK32B	JMJD1C	PSMC4	GPI
MYBL2	SIK3	GPI	PKMYT1	
CCNE2	HMGB3		PSMB2	
WDR62	ABCB6		RRM2	
E2F7	VPS4A		DLX2	
CENPA	PSMB2		DSN1	
TIMELESS	DLX2		PTTG1/PTTG2	
TK1	LYPD6		SAC3D1	
KIF20A	STC2		TROAP	
CKAP5	BNIP3L		TIMELESS	
C15ORF42	PTTG1/PTTG2		NUP93	



A, Survival curve for DE analysis

B, Survival curve for DN analysis

C, Survival curve for INDEED

A: Log rank test *p*-value = 0.0024
Hazard ratio = 2.75
Concordance index = 0.669

B: Log rank test *p*-value = 0.00065
Hazard ratio = 3.16
Concordance index = 0.678

C: Log rank test *p*-value = 5.64e-5
Hazard ratio = 4.12
Concordance index = 0.679

Figure 5.10 Survival curves for A) differential expression (DE) analysis , B) differential network (DN) analysis, and C) INDEED.

In summary, differential expression and differential network analyses identify biomarker candidates from two complementary perspectives: the former investigates the change of single biomolecule in its expression level between distinct biological groups, while the latter focuses on the change at the biomolecular pair level. The improved performance of INDEED is attributed to its capability to simultaneously consider the changes between cases and controls on individual biomolecule and bimolecular pair levels, while differential expression and differential network analyses can only capture changes on one of the two levels.

5.4 INDEED-M: extending INDEED for multi-omic data integration

One intuitive way to extend INDEED for multi-omic data integration is to concatenate different omic data on the same set of samples after proper normalization. Considering the Gaussian assumption of graphical LASSO, a normalization method that can preserve the ranking of all samples for each biomolecule, and simultaneously ensure the marginal probability distribution for each biomolecule is standard Gaussian distribution is desirable. Once the normalization is performed, we can apply INDEED on the concatenated data matrix. The resulting differential network will contain both intra-omic interactions that are connections within the same type of omic data (e.g., metabolite-metabolite) and inter-omic interactions that are connections among different types of omic data (metabolite-protein) as shown in Figure 1.1. The drawback of this intuitive method is that it treats each omic data type equally, while in real application, we usually prefer to focus on only one type of omic data (e.g., metabolomics) and use other types of omic data (e.g., proteomics and glycomics) to help discover more reliable and powerful biomarker candidates for the focused type of omic data (e.g., metabolomics) in a systems biology perspective. Without loss of generality, in Section 5.4, we would assume metabolomics is our focused data type, and proteomics and glycomics are available on the same set of samples for multi-omic data integration.

While intra-omic interactions can be measured through typical network modeling methods such as graphical LASSO in INDEED, studying inter-omic interactions is a challenge due to the disparate technologies used for generating different omic datasets. A few methods have been proposed to solve this problem including co-inertia analysis (CIA), sparse canonical correlation analysis (sCCA), sparse partial least square (sPLS), sparse generalized canonical correlation analysis (sGCCA), and multiple co-inertia analysis (MCIA) [91-94]. Most of these methods are complicated and computational expensive. Here, we will focus on the intra-omic interactions between metabolite pairs and use Spearman's correlation to investigate the inter-omic interactions between metabolomics data with other types of omic data (e.g., metabolite-protein, metabolite-glycan). Following the idea of INDEED, we will build an integrated differential network in which only rewiring intra-omic interactions (e.g., rewiring metabolite-metabolite pairs) and rewiring inter-omic interactions (e.g., rewiring metabolite-protein or metabolite-glycan pairs) exist. The rewiring intra-omic interactions will be obtained using INDEED on metabolomic dataset. The rewiring inter-omic interactions will be identified by computing Spearman's correlation between metabolites and other type of statistically significant biomolecules such as proteins and glycans. A similar activity score (Equation 5.7) will be calculated to prioritize the metabolite biomarker candidates to favor those that are connected with other types of statistically significant biomolecules (e.g., proteins and glycans) in the integrated differential network. The extended INDEED method is called INDEED-M. We believe metabolite biomarker candidates that are identified from the integrated differential network are more likely to be reliable biomarker candidates with the help of the multi-omic data. For example, proteins that are differentially expressed between biologically disparate groups, may serve as enzymes impacting metabolic reactions and cause the metabolites involved to be differentially expressed between the two groups as well. By extending INDEED on multi-omic data, we have a better chance to identify more reliable and powerful cancer biomarkers.

Figure 5.11 shows the extended INDEED framework (INDEED-M) for multi-omic data integration. Specifically, given multi-omic data with a focus on identifying metabolite biomarker candidates, we first apply INDEED on metabolomic dataset to identify rewiring metabolite-

metabolite intra-omic interactions. Then differential expression analysis is performed to identify statistically significant genes, proteins and glycans between distinct biological groups. Then Spearman's correlation will be used to calculate the intra-omic interactions between each metabolite and the statistically significant genes, protein and glycan in a group-specific manner ($c_{ij}^{(1)}$ and $c_{ij}^{(2)}$). The change for each inter-omic interaction between distinct biological groups is calculated as $\Delta c_{ij} = c_{ij}^{(1)} - c_{ij}^{(2)}$. To obtain the rewiring inter-omic interactions between the two groups, permutation test will be performed as similar to INDEED. First, it randomly permutes the order of the samples within a biological group for each gene, protein, metabolite, and glycan. Then, Spearman's correlation for each inter-omic interaction in each biological group is computed ($\tilde{c}_{ij}^{(1)}$ and $\tilde{c}_{ij}^{(2)}$). At last, the inter-omic interaction change between groups is computed ($\Delta \tilde{c}_{ij} = \tilde{c}_{ij}^{(1)} - \tilde{c}_{ij}^{(2)}$). This procedure is repeated 1000 times to obtain an empirical distribution of $\Delta \tilde{c}_{ij}$. $\Delta c_{ij} \neq 0$ is considered statistically significant if Δc_{ij} falls into the 2.5% tails on either end of the empirical distribution curve for $\Delta \tilde{c}_{ij}$. An integrated differential network, which contains genes, proteins, metabolites and glycans will be built by merging the rewiring intra-omic interactions (i.e., $\Delta pc_{ij} \neq 0$ is statistically significant) with the rewiring inter-omic interactions (i.e., $\Delta c_{ij} \neq 0$ is statistically significant). Finally, an activity score (s_k^M) will be calculated for each metabolite as the summation of z_k and the *z*-scores for all its neighbors in the integrated differential network. A higher activity score indicates that the corresponding metabolite has more rewiring interactions on both intra-omic and inter-omic levels with its neighbors, and their changes on intensity level have higher statistically significance levels between the groups.

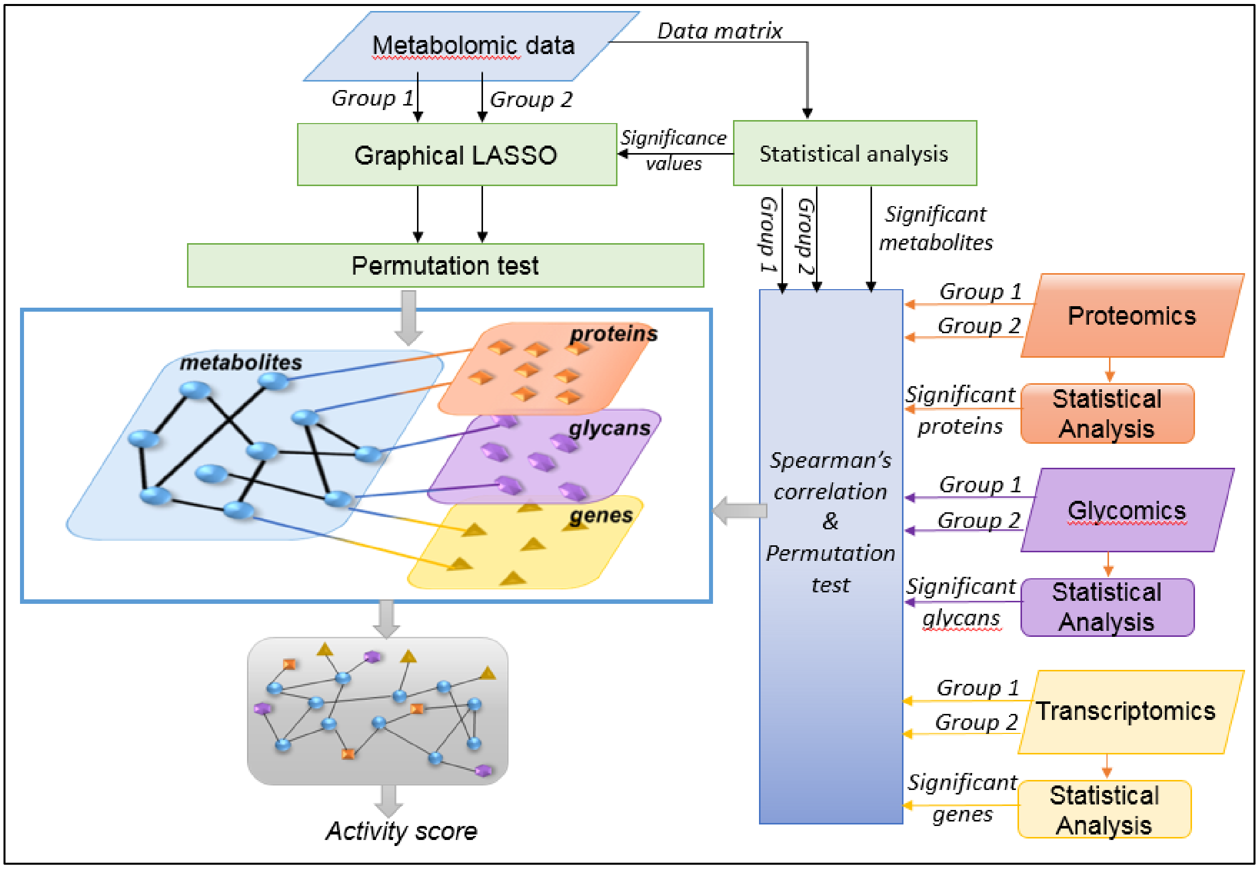


Figure 5.11 The framework of INDEED-M. In differential network analysis, the network is built based on correlation and partial correlation, using Spearman's correlation and graphical LASSO, respectively.

5.5 Evaluation of INDEED-M on multi-omic data

We evaluated INDEED-M on proteomic, metabolomics and glycomic data from the same set of samples. Similar to the proteomic and glycomic datasets used in Chapter 5.3.1 and 5.3.2, the metabolomics datasets were acquired by analysis of metabolites in sera from hepatocellular carcinoma (HCC) cases and liver cirrhotic controls [95]. Only TU cohort is used for multi-omic data integration. Briefly, adult patients were recruited from the outpatient clinics and inpatient wards of the Tanta University Hospital in Tanta, Egypt. The participants consist of 89 subjects (40 HCC cases and 49 patients with liver cirrhosis). We used gas chromatography coupled with mass spectrometry (GC-MS) for both untargeted and targeted analyses of sera from subjects in TU cohort. Metabolites that are statistically significant between the two groups were selected from the untargeted metabolomic data. A total of 40 metabolites were then evaluated by targeted analysis in the same sera samples, using GC-qMS in selected ion monitoring (SIM) mode. More details on experiment design and statistical analysis can be found in [95].

Figure 5.12 depicts a differential network built using INDEED from two group-specific networks (CIRR and HCC groups) based on 40 metabolites measured in sera from HCC cases and cirrhotic controls in the TU cohort. The edges represent significant rewiring intra-omic interactions between metabolite pairs based on the permutation test. Through activity score (Equation 5.7), we ranked the metabolite biomarker candidates for subsequent analysis. Furthermore, we integrated the metabolomic data with proteomic and glycomic data acquired on the same set of samples to investigate the rewiring inter-omic interactions (i.e., metabolite-protein and metabolite-glycan). Spearman’s correlation was calculated between each metabolite and protein/glycan in a group-specific manner and tested for significance through permutation. We merged the differential network in Figure 5.12 with Spearman’s correlation results to build an integrated differential network (Figure 5.13). Through visualization of Figures 5.12 and 5.13 and the defined activity score, we ranked the metabolite biomarker candidates.

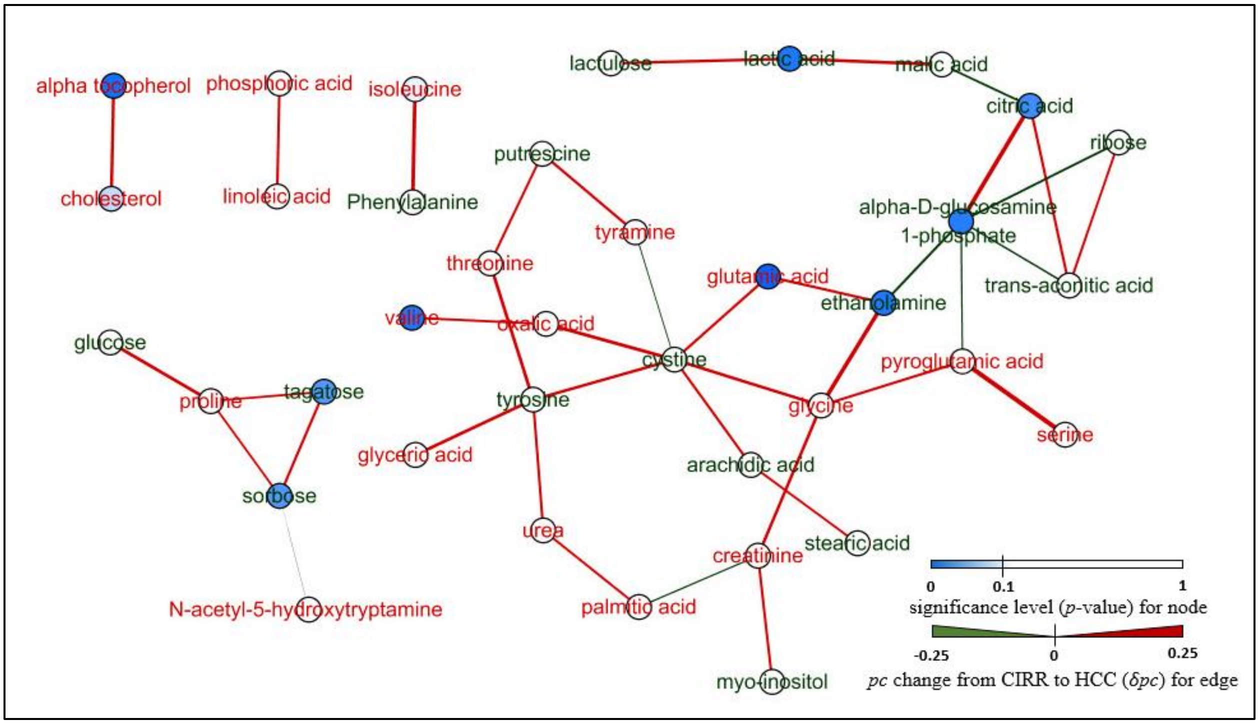


Figure 5.12 Differential networks built by INDEED based on metabolomic data. Node color indicates the significance level of the metabolite's change in HCC vs. CIRR. Red and green labels denote up and down regulation in HCC vs. cirrhosis, respectively. Edge color and thickness indicate change on partial correlation (pc) for metabolite pair from CIRR to HCC group ($\delta pc = pc_{HCC} - pc_{CIRR}$).

Table 5.6 presents 12 significant metabolites (p-value < 0.05) found by t-test [95], and the top 12 metabolites ranked by INDEED and INDEED-M methods. While the three lists of metabolites largely overlap, there are some differences. For example, malic acid, cystine, and proline were missed by t-test, but were ranked in the top 12 by both INDEED and INDEED-M. A literature survey reveals that these metabolites were reported in previous studies to have association with HCC, thus demonstrating the potential of using INDEED and INDEED-M as alternatives to typical differential expression analysis for cancer biomarker discovery [96-98].

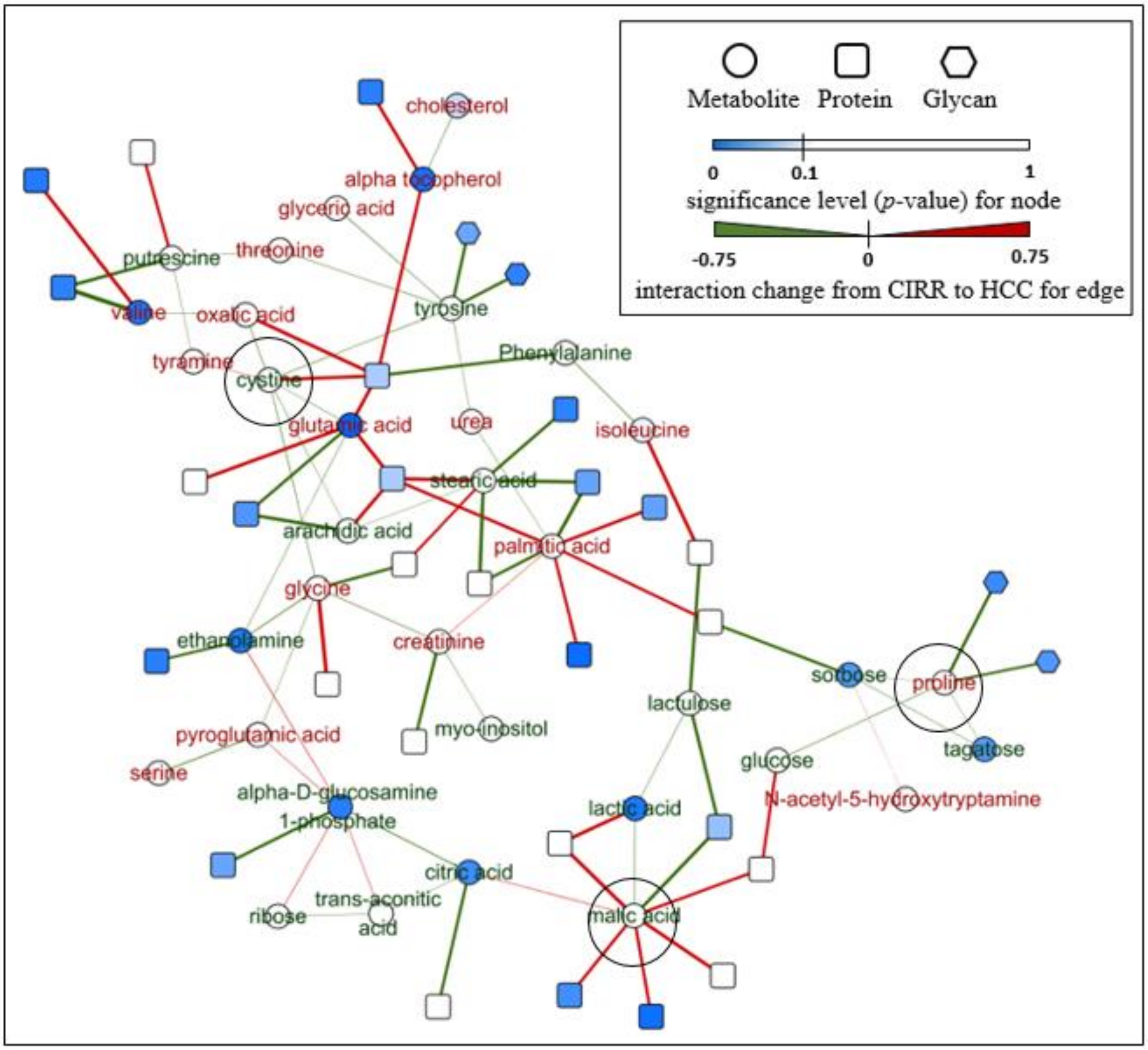


Figure 5.13 Integrative differential network derived from three omic datasets using INDEED-M. Metabolites are represented by circles, proteins (rectangles) and glycans (hexagons). Node color indicates the significance level of the biomolecule's change in HCC vs. CIRR. Red and green labels denote up and down regulation in HCC vs. cirrhosis, respectively. Edge color and thickness indicate change on Spearman's correlation for inter-omic interaction or partial correlation (*pc*) for intra-omic interaction of a biomolecular pair from CIRR to HCC group. Cystine, malic acid and proline are circled out.

Table 5.6 HCC-associated metabolites selected by t-test (p-value < 0.05) and ranked by the network-based differential analysis and multi-omic data integration. Cystine, malic acid and proline are in bold.

Selected by t-test (p-value < 0.05)	**p-value**	**Ranked by INDEED**	**Activity score**	**p-value**	**Ranked by INDEED-M**	**Activity score**	**p-value**
glutamic acid	5.5E-08	ethanolamine	9.7	0.007	**malic acid**	15.8	0.060
alpha tocopherol	0.002	alpha-D-glucosamine 1-phosphate	9.0	0.007	glutamic acid	14.1	5.5E-08
valine	0.003	**cystine**	8.4	0.488	palmitic acid	12.6	0/386
lactic acid	0.003	glutamic acid	6.9	5.5E-08	enthanolamine	12.1	0.007
ethanolamine	0.007	citric acid	6.9	0.011	alpha-D-glucosamine 1-phosphate	11.1	0.007
alpha-D-glucosamine 1-phosphate	0.007	trans-aconitic acid	6.5	0.254	**cystine**	10.2	0.488
citric acid	0.011	**malic acid**	6.4	0.060	**proline**	9.3	0.321
sorbose	0.015	sorbose	5.8	0.015	stearic acid	9.1	0.421
tagatose	0.015	lactic acid	5.4	0.003	valine	8.7	0.003
leucine	0.015	**proline**	4.9	0.321	alpha tocopherol	8.7	0.002
cholestrol	0.036	tagatose	4.7	0.015	citric acid	8.2	0.011
isoleucine	0.042	alpha tocopherol	4.4	0.002	tyrosine	7.5	0.336

5.6 Conclusion

In Chapter 5, we propose a novel approach, INDEED, to build a sparse differential network based on partial correlation for better visualization, and integrate differential expression and differential network analyses for biomarker discovery on single omic dataset. The application of INDEED on real transcriptomic, proteomic and glycomic data reveals its potential to select biomarker candidates that are more reproducible across independent studies, and leads to improved classification and regression accuracy when compared with differential expression and differential network analyses, separately. We further extend INDEED to INDEED-M for multi-omic data integration. The integrated differential network from INDEED-M leads to new biomarker candidates that are missed by typical differential expression analysis, but are relevant to the disease under study through real multi-omic data application. We also develop an open source R package to share INDEED and INDEED-M with the scientific community at Github (https://github.com/Hurricaner1989/INDEED-R-package).

6 Contribution and Future work

6.1 Contribution

6.1.1 Summary of contribution

In this dissertation, we investigate the potential of integrating differential expression and differential network analyses for biomarker discovery on single and multi-omic data. In order to achieve this goal, we have developed a series of novel methods to reconstruct sparse biological network, incorporate prior biological knowledge into data-driven network model and perform differential network analysis to identify cancer biomarker candidates. Briefly, these include:

- We developed a novel method LOPC to reconstruct sparse biological network using partial correlation.
- We developed a novel method wgLASSO to incorporate prior biological knowledge into data-driven network model to build more biologically relevant network.
- We developed a novel differential network analysis method dwgLASSO to identify cancer biomarker candidates based on the group-specific networks built by wgLASSO.
- We developed a novel method INDEED to integrate differential expression and differential network analyses on single omic data to identify cancer biomarker candidates that have significant changes on both single biomolecule and biomolecular pair levels.
- We extended INDEED to INDEED-M for integrative analysis of multi-omic data to identify cancer biomarker candidates by their changes on both single biomolecule and biomolecular pair levels at different layers of human biological system (e.g., genes, proteins, metabolites).

- We developed Matlab and R packages to implement LOPC, wgLASSO, dwgLASSO, INDEED and INDEED-M, and uploaded them at Github (https://github.com/Hurricaner1989) to share with the research community.

6.1.2 LOPC

LOPC is proposed to reconstruct sparse biological networks by calculating up to the second order partial correlation. Simulation results show that, compared with other undirected network inference methods (correlation, GGM, and 0-1 graph), LOPC has better performance in reconstructing networks with less spurious edges (false positives). It also works well under the 'large *p* small *n*' restriction. These properties make LOPC a promising network model to reconstruct sparse biological networks from omic datasets, which may give insights into the mechanisms of complex diseases. A real application of LOPC on metabolomics dataset validates its performance and shows its potential in discovering biomarker candidates that are missed by typical statistical analysis methods such as t-test and ANOVA. An open source Matlab package is available at Github to share LOPC with the scientific community (https://github.com/Hurricaner1989/LOPC-Matlab-package).

6.1.3 wgLASSO and dwgLASSO

A novel network reconstruction method, wgLASSO is proposed to integrate prior biological knowledge into data-driven model. Simulation results show that wgLASSO can achieve better performance in reconstructing biologically networks than purely data-driven network models (e.g., neighbor selection and graphical LASSO) even when only a moderate level of information is available as prior biological knowledge. We further develop a novel differential network analysis method dwgLASSO based on the group specific networks built by wgLASSO. We evaluate the performance of dwgLASSO in survival time prediction using two independent microarray breast cancer datasets previously published by Bild *et al.* and van de Vijver *et al.* The top 10 ranking genes selected by dwgLASSO based on the dataset from Bild *et al.* lead to a significantly improved survival prediction on the independent dataset from van de Vijver *et al.*, compared with the top 10 significant genes obtained by conventional differential gene expression analysis. Among the top 10 genes selected by dwgLASSO, UBE2S, SALL2, XBP1 and KIAA0922 have been previously reported to be relevant in breast cancer biomarker discovery study. We also test dwgLASSO using TCGA RNA-seq data acquired from patients with HCC on tumors samples and their corresponding non-tumorous liver tissues. Improved sensitivity, specificity and AUC are observed when comparing dwgLASSO with conventional differential gene expression analysis method. Both wgLASSO and dwgLASSO are available as an open source R package at Github to share with the scientific community (https://github.com/Hurricaner1989/dwgLASSO-R-package).

6.1.4 INDEED and INDEED-M

A novel approach, INDEED, is proposed to integrate differential expression and differential network analyses for biomarker discovery on single omic dataset. The application of INDEED on real transcriptomic, proteomic and glycomic data reveals its potential to select biomarker candidates that are more reproducible across independent studies, and leads to improved classification and regression accuracy when compared with differential expression and differential network analyses, separately. We further extend INDEED to INDEED-M for integrative analysis of multi-omic data. The integrated differential network from INDEED-M leads to new biomarker candidates that are missed by typical differential expression analysis, but are relevant to the disease under study in real multi-omic data application. We also develop an open source R package to share INDEED and INDEED-M with the scientific community at Github (https://github.com/Hurricaner1989/INDEED-R-package).

6.2 Future work

Future work involves some further development of the proposed methods. To be more specific, for LOPC, the computational burden to reconstruct a sparse biological network that contains thousands of biomolecules is heavy due to its requirement to calculate up to the second order partial correlation. Further work that can simplify the calculation of the second order partial correlation will reduce the computational time significantly. In addition, an appropriate approach to integrate prior biological knowledge into LOPC will also be desirable.

For wgLASSO and dwgLASSO, further development will be investigating the potential of using multiple databases simultaneously to increase the coverage of biomolecular pairs that have prior biological knowledge. In addition, the differential network score defined within dwgLASSO can be further modified to calculate more accurately the rewiring interactions between biological distinct groups for a given biomolecule.

For INDEED and INDEED-M, the activity score is defined as the summation of the z-score of all its neighbors of one node, including itself. As a result, the prioritized list will favor biomarker candidates that have more rewiring connections over small p-values. Further modification on the activity score should balance the trade-off between rewiring connections and p-values in a more reasonable way. Additionally, in INDEED-M, Spearman's correlation might be replaced by other more advanced measurements to better characterize the inter-omic interactions between different types of biomolecules. At last, it will be desirable to integrate prior biological knowledge into INDEED and INDEED-M in building differential network to indicate which rewiring connections have supporting evidence from public available databases for better explanation of the identified biomarker candidates.

Appendix A. Proof of the equivalence of partial correlation and conditional correlation under Gaussian distribution.

Conditional independence is crucial for network inference. Here, we prove the equality of partial correlation and conditional correlation involving three variables under Gaussian assumption so that we can use partial correlation to infer conditional independence relationships between nodes and build a network.

By definition, the partial correlation coefficient between x and y conditional on z $(r_{xy \cdot z})$ is obtained by first regressing x on z and y on z separately and then calculating the correlation between the residuals of the models for x and y:

$$r_{xy \cdot z} = corr(\hat{\varepsilon}_x,\ \hat{\varepsilon}_y) \qquad \text{(A-1)}$$

$$\hat{\varepsilon}_x = x - \hat{a} - \hat{b} \times z \qquad \text{(A-2)}$$

$$\hat{\varepsilon}_y = y - \hat{c} - \hat{d} \times z \qquad \text{(A-3)}$$

where $\hat{\varepsilon}_x$, $\hat{\varepsilon}_y$ are the residuals of x and y after regressing on z; $\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{d}$ are regression coefficients.

Conditional correlation coefficient between x and y given z ($r_{xy|z}$) is defined as:

$$r_{xy|z} = \frac{E_{xy|z}\{[x-E(X|Z)][y-E(y|Z)]\}}{\sqrt{E_{x|z}\{[x-E(X|Z)]^2\}E_{y|z}\left\{[y-E(y|Z)]^2\right\}}} \qquad \text{(A-4)}$$

where $E_{x|z}$, $E_{y|z}$ and $E_{x,y|z}$ denote expectations of the marginal and joint distribution of x and y conditional on z.

To show the relationship between partial correlation and conditional correlation, we consider the following case of $x = b_0 + b_1 \times z + a$ and $y = d_0 + d_1 \times z + c$, where b_0, b_1, d_0 and d_1 are constants, x, y, z, a and c are random variables. Under this assumption, the conditional correlation between x and y given z is reduced to:

$$r_{xy|z} = \frac{E_{x,y|z}\{[a-E(a)][c-E(c)]\}}{\sqrt{E_{x|z}\{[a-E(a)]^2\}E_{y|z}\{[c-E(c)]^2\}}} = \frac{Cov(a,c)}{\sqrt{Var(a)Var(c)}} = corr(a,c) = r_{xy\cdot z} \qquad \text{(A-5)}$$

From Eq. (A-5), the partial correlation equals to conditional correlation. To be more general, these two correlations are the same when the conditional variance and covariance of x and y given z are free of z [26, 27]. The above condition is satisfied in normal distribution.

Appendix B. Personal information

Biography

Yiming Zuo received his B.S. degree in the department of Information Science & Electronic Engineering from Zhejiang University, China in 2012. Currently, he is a fifth-year Ph.D. candidate in the department of Electrical & Computer Engineering at Virginia Polytechnic Institute and State University (Virginia Tech) under the supervision of Dr. Guoqiang Yu (Virginia Tech) and Dr. Habtom W. Ressom (Georgetown U.). His research interests are applying machine learning technique and probabilistic graphical model to reconstruct biological networks and integrate multi-omic data for cancer biomarker discovery.

List of relevant publications

[1], **Zuo Y.**, Cui Y., Yu G., Li R., & Ressom H. W. (2017). Incorporating prior knowledge for network-based differential gene expression analysis using weighted graphical LASSO. BMC Bioinformatics, 18(1), 99.

[2], Di Poto C., Ferrarini A., Zhao Y., Varghese R., Tu C., **Zuo Y.**, Wang M., Ranjbar M. R. N., Luo Y., Zhang C., Desai C. S., Shetty K., Tadesse M. G., & Ressom H. W. (2016). Metabolomic Characterization of Hepatocellular Carcinoma in Patients with Liver Cirrhosis for Biomarker Discovery. Cancer Epidemiology and Prevention Biomarkers, cebp-0366.

[3], **Zuo, Y.**, Cui, Y., Di Poto, C., Varghese, R. S., Yu, G., Li, R., & Ressom, H. W. (2016). INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery. Methods, 111, 12-20.

[4], Wang, J., **Zuo, Y.**, Man, Y. G., Avital, I., Stojadinovic, A., Liu, M., ... & Ressom, H. W. (2015). Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. Journal of Cancer, 6(1), 54-65.

[5], **Zuo Y.**, Yu G., Tadesse M. G., & Ressom H. W. (2014). Biological network inference using low order partial correlation. Methods, 69(3), 266-273.

[6], Wang, J., **Zuo, Y.**, Liu, L., Man, Y., Tadesse, M. G., & Ressom, H. W. (2014). Identification of functional modules by integration of multiple data sources using a bayesian network classifier, Circulation: Cardiovascular Genetics, 7(2), 206-217.

[7], Varghese R. S., **Zuo Y.**, Zhao Y., Zhang Y., Jablonski S. A., Pierobon M., Petricoin E. F., Ressom H. W., Weiner L. M. (2016). Network-Based Analysis of Reverse Phase Protein Array Data. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 312-317). IEEE.

[8], Ressom H. W., Di Poto C., Ferrarini A., Hu Y., Ranjbar M. N., Song E., Varghese R., Wang M., Zhou S., Zhu R., **Zuo Y.**, & Tadesse M., Mechref Y. (2016). Multi-Omic Approaches for Characterization of Hepatocellular Carcinoma. In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the (pp. 3437-3440). IEEE.

[9], **Zuo Y.**, Yu G., & Ressom H. W. (2015). Integrating prior biological knowledge and graphical LASSO for network inference. In Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on (pp. 1543-1547). IEEE.

[10], **Zuo Y.**, Yu G., Zhang C., & Ressom H. W. (2014). A new approach for multi-omic data integration. In Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on (pp. 214-217). IEEE.

[11], **Zuo Y.**, Yu G., Tadesse M. G., & Ressom H. W. (2013). Reconstructing biological networks using low order partial correlation. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on (pp. 171-175). IEEE.

Bibliography

[1] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences,* vol. 103, pp. 5923-5928, 2006.

[2] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette*, et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences,* vol. 102, pp. 15545-15550, 2005.

[3] Y. Zuo, G. Yu, C. Zhang, and H. W. Ressom, "A new approach for multi-omic data integration," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 2014, pp. 214-217.

[4] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network ‐ based classification of breast cancer metastasis," *Molecular systems biology,* vol. 3, 2007.

[5] A. de la Fuente, "From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases," *Trends in genetics,* vol. 26, pp. 326-333, 2010.

[6] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics,* vol. 12, pp. 56-68, 2011.

[7] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular systems biology,* vol. 8, p. 565, 2012.

[8] Y. Zuo, G. Yu, M. G. Tadesse, and H. W. Ressom, "Biological network inference using low order partial correlation," *Methods,* vol. 69, pp. 266-273, 2014.

[9] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen, "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene," *Nucleic acids research,* vol. 28, pp. 3442-3444, 2000.

[10] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research,* vol. 34, pp. D535-D539, 2006.

[11] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono*, et al.*, "Reactome: a knowledgebase of biological pathways," *Nucleic acids research,* vol. 33, pp. D428-D432, 2005.

[12] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, "ConsensusPathDB—a database for integrating human functional interaction networks," *Nucleic acids research,* vol. 37, pp. D623-D628, 2009.

[13] A. Reverter, A. Ingham, S. A. Lehnert, S.-H. Tan, Y. Wang, A. Ratnakumar, *et al.,* "Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer," *Bioinformatics,* vol. 22, pp. 2396-2404, 2006.

[14] C. Y. Zuo Y., Yu G., Li R., & Ressom H. W, "Incorporating prior knowledge for network-based differential gene expression analysis using weighted graphical LASSO.," *BMC Bioinformatics,* 2016.

[15] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of computational biology,* vol. 7, pp. 601-620, 2000.

[16] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics,* vol. 8, pp. 450-461, 2007.

[17] A. J. Butte and I. S. Kohane, "Unsupervised knowledge discovery in medical databases using relevance networks," in *Proceedings of the AMIA Symposium*, 1999, p. 711.

[18] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput,* 2000, pp. 418-429.

[19] J. Schafer and K. Strimmer, "Learning large-scale graphical Gaussian models from genomic data," *Science of Complex Networks from Biology to the Internet and WWW,* vol. 776, pp. 263-276, 2005.

[20] A. De La Fuente, N. Bing, I. Hoeschele, and P. Mendes, "Discovery of meaningful associations in genomic data using partial correlation coefficients," *Bioinformatics,* vol. 20, pp. 3565-3574, 2004.

[21] P. M. Magwene and J. Kim, "Estimating genomic coexpression networks using first-order conditional independence," *Genome Biol,* vol. 5, p. R100, 2004.

[22] A. Wille and P. Bühlmann, "Low-order conditional independence graphs for inferring genetic networks," *Statistical applications in genetics and molecular biology,* vol. 5, 2006.

[23] A. Reverter and E. K. Chan, "Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks," *Bioinformatics,* vol. 24, pp. 2491-2497, 2008.

[24] L. M. De Campos and J. F. Huete, "A new approach for learning belief networks using independence criteria," *International Journal of Approximate Reasoning,* vol. 24, pp. 11-37, 2000.

[25] S. Lauritzen, "Appendix B. linear algebra and random vectors," *Graphical Models,* pp. 243-244, 1996.

[26] A. Lawrance, "On conditional and partial correlation," *The American Statistician,* vol. 30, pp. 146-149, 1976.

[27] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics,* vol. 46, pp. 657-664, 2004.

[28] T. Anderson, "2.5. 3. some formulas for partial correlations,"" *An Introduction to Multivariate Statistical Analysis,* pp. 39-41, 2003.

[29] M. E. Olobatuyi, *A user's guide to path analysis*: University Press of America, 2006.

[30] R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika,* vol. 10, pp. 507-521, 1915.

[31] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature,* vol. 407, pp. 651-654, 2000.

[32] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science,* vol. 296, pp. 910-913, 2002.

[33] D. Edwards, *Introduction to graphical modelling*: Springer Science & Business Media, 2012.

[34] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics,* vol. 9, pp. 432-441, 2008.

[35] J. F. Xiao, R. S. Varghese, B. Zhou, M. R. Nezami Ranjbar, Y. Zhao, T.-H. Tsai, *et al.,* "LC–MS based serum metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort," *Journal of proteome research,* vol. 11, pp. 5914-5923, 2012.

[36] H. Toh and K. Horimoto, "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling," *Bioinformatics,* vol. 18, pp. 287-297, 2002.

[37] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis,* vol. 90, pp. 196-212, 2004.

[38] H. Kishino and P. J. Waddell, "Correspondence analysis of genes and tissue types and finding genetic links from microarray data," *Genome Informatics,* vol. 11, pp. 83-95, 2000.

[39] A. P. Dempster, "Covariance selection," *Biometrics,* pp. 157-175, 1972.

[40] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics,* vol. 21, pp. 754-764, 2005.

[41] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics,* pp. 1436-1462, 2006.

[42] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electronic journal of statistics,* vol. 6, p. 2125, 2012.

[43] Y. Zuo, G. Yu, and H. W. Ressom, "Integrating prior biological knowledge and graphical LASSO for network inference," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on,* 2015, pp. 1543-1547.

[44] Z. Wang, W. Xu, F. A. San Lucas, and Y. Liu, "Incorporating prior knowledge into gene network study," *Bioinformatics,* vol. 29, pp. 2633-2640, 2013.

[45] Y. Li and S. A. Jackson, "Gene Network Reconstruction by Integration of Prior Biological Knowledge," *G3: Genes| Genomes| Genetics,* vol. 5, pp. 1075-1079, 2015.

[46] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288, 1996.

[47] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics,* vol. 26, pp. 976-978, 2010.

[48] H. Caniza, A. E. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, *et al.*, "GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology," *Bioinformatics,* vol. 30, pp. 2235-2236, 2014.

[49] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 72, pp. 417-473, 2010.

[50] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews genetics,* vol. 5, pp. 101-113, 2004.

[51] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman, "The huge package for high-dimensional undirected graph estimation in R," *The Journal of Machine Learning Research,* vol. 13, pp. 1059-1062, 2012.

[52] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 5116-5121, 2001.

[53] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K.-W. Tsui, "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *Journal of computational biology,* vol. 8, pp. 37-52, 2001.

[54] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayes analysis of a microarray experiment," *Journal of the American statistical association,* vol. 96, pp. 1151-1160, 2001.

[55] J. E. Hutz, A. T. Kraja, H. L. McLeod, and M. A. Province, "CANDID: a flexible method for prioritizing candidate genes for complex human traits," *Genetic epidemiology,* vol. 32, p. 779, 2008.

[56] L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens*, et al.*, "ENDEAVOUR update: a web resource for gene prioritization in multiple species," *Nucleic acids research,* vol. 36, pp. W377-W384, 2008.

[57] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics,* vol. 82, pp. 949-958, 2008.

[58] L. Franke, H. Van Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *The American Journal of Human Genetics,* vol. 78, pp. 1011-1025, 2006.

[59] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle*, et al.*, "An integrated approach to inferring gene–disease associations in humans," *Proteins: Structure, Function, and Bioinformatics,* vol. 72, pp. 1030-1037, 2008.

[60] M. J. Ha, V. Baladandayuthapani, and K.-A. Do, "DINGO: differential network analysis in genomics," *Bioinformatics,* vol. 31, pp. 3413-3420, 2015.

[61] B. Zhang, H. Li, R. B. Riggins, M. Zhan, J. Xuan, Z. Zhang*, et al.*, "Differential dependency network analysis to identify condition-specific topological changes in biological networks," *Bioinformatics,* vol. 25, pp. 526-532, 2009.

[62] Y. Tian, B. Zhang, E. P. Hoffman, R. Clarke, Z. Zhang, I.-M. Shih*, et al.*, "Knowledge-fused differential dependency network models for detecting significant rewiring in biological networks," *BMC systems biology,* vol. 8, p. 1, 2014.

[63] Y. Tian, B. Zhang, E. P. Hoffman, R. Clarke, Z. Zhang, I.-M. Shih*, et al.*, "KDDN: an open-source Cytoscape app for constructing differential dependency networks with significant rewiring," *Bioinformatics,* vol. 31, pp. 287-289, 2015.

[64] Z. Wei and H. Li, "A Markov random field model for network-based analysis of genomic data," *Bioinformatics,* vol. 23, pp. 1537-1544, 2007.

[65] P. Chouvardas, G. Kollias, and C. Nikolaou, "Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis," *BMC bioinformatics,* vol. 17, p. 181, 2016.

[66] P. Wei and W. Pan, "Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model," *Bioinformatics,* vol. 24, pp. 404-411, 2008.

[67] H. Binder and M. Schumacher, "Incorporating pathway information into boosting estimation of high-dimensional risk prediction models," *BMC bioinformatics,* vol. 10, p. 18, 2009.

[68] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse*, et al.*, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature,* vol. 439, pp. 353-357, 2006.

[69] M. J. Van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil*, et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine,* vol. 347, pp. 1999-2009, 2002.

[70] A. J. Gentles, A. M. Newman, C. L. Liu, S. V. Bratman, W. Feng, D. Kim*, et al.*, "The prognostic landscape of genes and infiltrating immune cells across human cancers," *Nature medicine,* vol. 21, pp. 938-945, 2015.

[71] M. J. Pencina and R. B. D'Agostino, "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation," *Statistics in medicine,* vol. 23, pp. 2109-2123, 2004.

[72] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom, "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO," *BMC bioinformatics,* vol. 18, p. 99, 2017.

[73] A. K. Ayesha, T. Hyodo, E. Asano, N. Sato, M. A. Mansour, S. Ito*, et al.*, "UBE2S is associated with malignant characteristics of breast cancer cells," *Tumor Biology,* pp. 1-10, 2015.

[74] E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng*, et al.*, "Gene expression predictors of breast cancer outcomes," *The Lancet,* vol. 361, pp. 1590-1596, 2003.

[75] H. Liu, A. S. Adler, E. Segal, and H. Y. Chang, "A transcriptional program mediating entry into cellular quiescence," *PLoS Genet,* vol. 3, p. e91, 2007.

[76] X. Chen, D. Iliopoulos, Q. Zhang, Q. Tang, M. B. Greenblatt, M. Hatziapostolou*, et al.*, "XBP1 promotes triple-negative breast cancer by controlling the HIF1 [agr] pathway," *Nature,* vol. 508, pp. 103-107, 2014.

[77] N. Maharzi, V. Parietti, E. Nelson, S. Denti, M. Robledo-Sarmiento, N. Setterblad*, et al.*, "Identification of TMEM131L as a novel regulator of thymocyte proliferation in humans," *The Journal of Immunology,* vol. 190, pp. 6187-6197, 2013.

[78] J. Zhu, J. Z. Sanborn, S. Benz, C. Szeto, F. Hsu, R. M. Kuhn*, et al.*, "The UCSC cancer genomics browser," *Nature methods,* vol. 6, pp. 239-240, 2009.

[79] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology,* vol. 15, p. 550, 2014.

[80] M. Padi and J. Quackenbush, "Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators," *BMC systems biology,* vol. 9, p. 1, 2015.

[81] T. H. Tsai, E. Song, R. Zhu, C. Di Poto, M. Wang, Y. Luo*, et al.*, "LC‐MS/MS‐based serum proteomics for identification of candidate biomarkers for hepatocellular carcinoma," *Proteomics,* vol. 15, pp. 2369-2381, 2015.

[82] T.-H. Tsai, M. Wang, C. Di Poto, Y. Hu, S. Zhou, Y. Zhao*, et al.*, "LC–MS profiling of N-glycans derived from human serum samples for biomarker discovery in hepatocellular carcinoma," *Journal of proteome research,* vol. 13, pp. 4859-4868, 2014.

[83] Y. Pawitan, J. Bjöhle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall*, et al.*, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast Cancer Research,* vol. 7, p. 1, 2005.

[84] D. M. Witten, J. H. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *Journal of Computational and Graphical Statistics,* vol. 20, pp. 892-900, 2011.

[85] Y. Zuo, Y. Cui, C. Di Poto, R. S. Varghese, G. Yu, R. Li*, et al.*, "INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery," *Methods,* vol. 111, pp. 12-20, 2016.

[86] C. C. Liu, Y. H. Wang, E. Y. Chuang, M. H. Tsai, Y. H. Chuang, C. L. Lin*, et al.*, "Identification of a liver cirrhosis signature in plasma for predicting hepatocellular carcinoma risk in a population‐based cohort of hepatitis B carriers," *Molecular carcinogenesis,* vol. 53, pp. 58-66, 2014.

[87] X. He, Y. Wang, W. Zhang, H. Li, R. Luo, Y. Zhou*, et al.*, "Screening differential expression of serum proteins in AFP-negative HBV-related hepatocellular carcinoma using iTRAQ-MALDI-MS/MS," *Neoplasma,* vol. 61, pp. 17-26, 2013.

[88] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software,* vol. 33, p. 1, 2010.

[89] A. J. Gentles, A. M. Newman, C. L. Liu, S. V. Bratman, W. Feng, D. Kim*, et al.*, "The prognostic landscape of genes and infiltrating immune cells across human cancers," *Nat Med,* vol. 21, pp. 938-45, Aug 2015.

[90] L. Marchionni, B. Afsari, D. Geman, and J. T. Leek, "A simple and reproducible breast cancer prognostic test," *BMC genomics,* vol. 14, p. 1, 2013.

[91] A. C. Culhane, G. Perrière, and D. G. Higgins, "Cross-platform comparison and visualisation of gene expression data using co-inertia analysis," *BMC bioinformatics,* vol. 4, p. 59, 2003.

[92] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology,* vol. 8, pp. 1-34, 2009.

[93] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill, and V. Frouin, "Variable selection for generalized canonical correlation analysis," *Biostatistics,* p. kxu001, 2014.

[94] C. Meng, B. Kuster, A. C. Culhane, and A. M. Gholami, "A multivariate approach to the integration of multi-omics datasets," *BMC bioinformatics,* vol. 15, p. 162, 2014.

[95] M. R. N. Ranjbar, Y. Luo, C. Di Poto, R. S. Varghese, A. Ferrarini, C. Zhang*, et al.*, "GC-MS based plasma metabolomics for identification of candidate biomarkers for hepatocellular carcinoma in Egyptian cohort," *PloS one,* vol. 10, p. e0127299, 2015.

[96] Q. Huang, Y. Tan, P. Yin, G. Ye, P. Gao, X. Lu*, et al.*, "Metabolic characterization of hepatocellular carcinoma using nontargeted tissue metabolomics," *Cancer research,* vol. 73, pp. 4992-5002, 2013.

[97] M. Yu, Y. Zhu, Q. Cong, and C. Wu, "Metabonomics Research Progress on Liver Diseases," *Canadian Journal of Gastroenterology and Hepatology,* vol. 2017, 2017.

[98] T. Nishizaki, T. Matsumata, A. Taketomi, K. Yamamoto, and K. Sugimachi, "Levels of amino acids in human hepatocellular carcinoma and adjacent liver tissue," 1995.