IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 28, 2020, accepted June 3, 2020, date of publication June 5, 2020, date of current version June 17, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3000432*

# A Novel Network Modelling for Metabolite Set Analysis: A Case Study on CRC Metabolomics

**YUEYUE LIU[1], XIANGNAN XU[2], LINGLI DENG[3], KIAN-KAI CHENG[4], JINGJING XU[1], DANIEL RAFTERY[5], AND JIYANG DONG[1]**

[1]Department of Electronic Science, Xiamen University, Xiamen 361005, China
[2]School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia
[3]School of Information Engineering, East China University of Technology, Nanchang 330013, China
[4]Innovation Centre in Agritechnology, Universiti Teknologi Malaysia, Johor 84600, Malaysia
[5]Northwest Metabolomics Research Center, Anesthesiology and Pain Medicine, University of Washington, Seattle, WA 98109, USA

Corresponding author: Jiyang Dong (jydong@xmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 81871445 and Grant 81801788, in part by the Cancer Consortium Support Grant under Grant 5 P30 CA015704, in part by the University of Washington, and in part by the Educational Commission of Jiangxi Province of China under Grant GJJ160591. The work of Kian-Kai Cheng was supported by the Research University Grant from Universiti Teknologi Malaysia under Grant Q.J130000.2509.18H73.

**ABSTRACT** In metabolomics, pathway analysis normally refers to analysis of a pre-defined sets of metabolites (metabolite set) associated to the *metabolic* pathways. The metabolite set analysis is useful to facilitate biological interpretation of metabolomics data. The currently available methods may be divided into three generations: over-representation analysis, functional class scoring, and network topology analysis. Among the three generations of tools, the network topology methods have been shown to have lower false discovery rates and better biological interpretability than the other two earlier generations of tools. However, most of the current network topology methods focus the analysis only at the metabolite-level network. The interaction between pathways are not taken into consideration. To address this issue, we propose a new metabolite sets association network (MSAN) modelling scheme. In the developed method, the metabolite sets are defined based on the KEGG databases. By using the metabolite sets as vertexes, the MSAN network evaluated the relationships between pairs of metabolite sets based on their mutual information. The impact of a single metabolite set on the overall network was evaluated by the MSAN network, which may help to uncover differential metabolite sets relevant to the underlying biology mechanism of the study. A metabolomic dataset from a published colorectal cancer (CRC) study is used to evaluate the performance of MSAN network to identify perturbed metabolite sets in colorectal cancer patients. The current results are compared to that of two commonly used methods, NetGSA and MetaboAnalyst, which are based on the metabolite-level network approach. The current method highlights a number of metabolites sets consistent with recent published CRC reports. Taken together, the proposed method may provide an alternative tool for the identification of dysregulated pathways and facilitate biological interpretation of metabolomics data.

**INDEX TERMS** Metabolite sets association network (MSAN), colorectal cancer (CRC) metabolomics, mutual information, pathway analysis.

## I. INTRODUCTION

Pathway analysis has become an useful tool in the field of metabolomics, as it provides functional insights into the roles of differential metabolites in the development and treatment of numerous diseases. Over the past two decades, more than a dozen of pathway analysis methods have been developed [1]–[10]. Ma and colleagues [9] divide these methods into three generations: over-representation analysis (ORA) [11]–[13], functional class scoring (FCS) [14], [15], and network topology analysis [16], [17]. Generally, each pathway can be regarded to have a predefined metabolite subset and reaction subset. The methods based on ORA and FCS mainly focus on the analysis of the metabolite subset, while the third-generation network topology methods take the two subsets into consideration in pathway analysis.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

Pathways are interconnected and not independent, and dysregulated pathways can alter the properties of many other pathways within an organism through interactions between pathways [18]–[20]. On the basis of ORA and FCS, network topology analysis describes the relationship of metabolites in the form of a network, and takes into account the topological structure information between metabolites to evaluate the importance of a certain pathway [9], [10]. Since the topological information is useful to explore the interrelationships between pathways, network-based analysis methods provide some advantages in uncovering the biological functions of the perturbed pathway.

Network construction and topology analysis are two key steps in network-based methods. There are two types of networks based on the method of network construction. The first type uses chemical reaction relationships between metabolites to build a network: If two metabolites are the substrate and product of a chemical reaction, then an edge is linked between them. As the connection relationship of the network has practical biochemical significance, the results based on this network will be more biologically meaningful and more interpretable.

Due to the limitations in the analytical instrument platforms, the number of detected metabolites is usually far less than the total number of known metabolites in the pathways. Based on the low number of metabolites, it is difficult to establish a connected network and to perform topology analysis. Therefore, another set of network-based methods is used that take advantage of the relationships among detected metabolite abundances, correlations, partial correlations, mutual information, etc. to build a metabolite level association network [1]–[3]. Then, the topology analysis is carried out on the association network.

The current network-based pathway analysis method is mainly based on the metabolite-level network (MLAN). Since a metabolic pathway (a predefined metabolite set) is a sub-network of MLAN network, it is difficult to quantify the impact of a subnetwork to the overall network. In addition, metabolic pathways are known to overlap with each other as most of the pathways share some common metabolites. Thus, it is not straightforward to evaluate the impact of each pathway on the overall network and the interaction among the pathways.

In information theory, mutual information (MI) is commonly used to measure the association between two random variables. MI does not assume any property of the dependence between two variables; thus, it is more general than linear measures such as the correlation coefficient, and is able to detect more interactions [21]. The concept of MI is widely used to infer interaction networks in different fields including chemical, biological, and social area. In the current paper, we propose a novel network modeling scheme for the construction of a metabolite set association network (MSAN). In the MSAN method, metabolite sets are constructed based on the metabolic pathways in the KEGG or HumanCyc databases. In the method, MI is used to estimate the association between two pathways. In addition, the data process inequality (DPI) principle is used to create sparsity from the MI matrix data, and collect it into an adjacent matrix of strong connected networks. Furthermore, linear dimension reduction method is applied to transform the data submatrix of pathways into a one-dimensional vector, which make it possible to calculate the MI value between two multivariate matrices of small sample size. In the proposed method, network robustness and pathway sensitivity are evaluated using the random walk with restart (RWR) process by comparing with that of the simulated random networks. The key differential metabolite sets are evaluated on the differential MSAN network of two different biological states. The developed method is applied to a published colorectal cancer (CRC) metabolomic dataset, and the results highlight a number of perturbed pathways consistent with the published reports.

## II. METHODS

### A. MUTUAL INFORMATION AND DATA PROCESSING INEQUALITY

Let $X$ and $Y$ be two random variables. Mutual information (MI) between $X$ and $Y$ is defined as follows:

$$I\left(X,Y\right)=H\left(X\right)+H\left(Y\right)-H(X,Y) \tag{1}$$

where $H\left(X\right)$ is the entropy of $X$. If the possible value of $X$ is discreet, i.e. $X\in\{x_1,\ldots,x_n\}$, with corresponding probability as $p(x_i)$, $H\left(X\right)$ can be calculated by

$$H\left(X\right)=-\sum\nolimits_i p(x_i)\log p(x_i) \tag{2}$$

Let $p\left(x_i,y_j\right)=\mathrm{Prob}(X=x_i,Y=y_i)$ be the joint probability, and $p(x_i)$ and $p(y_j)$ as marginal probability of $X$ and $Y$, then the MI between $X$ and $Y$ can be calculated as reported in [22],

$$I\left(X,Y\right)=\sum\nolimits_{i,j}p\left(x_i,y_j\right)\log\left(\frac{p\left(x_i,y_j\right)}{p\left(x_i\right)p\left(y_j\right)}\right) \tag{3}$$

Estimation of three probability density functions (PDFs), i.e., $p\left(x_i\right)$, $p\left(y_j\right)$ and $p\left(x_i,y_j\right)$ are required for the calculation of MI using (3). In general, the analytical formulae of the three PDFs are unknown, but kernel density estimation can be used to provide an efficient and robust estimation of PDFs for datasets with small sample size.

For any random variable $X$, the formula of kernel density estimation (KDE) is given by [23]

$$\hat{p}_h\left(X\right)=\frac{1}{n}\sum\nolimits_{i=1}^{n}K\left(\frac{X-x_i}{h}\right) \tag{4}$$

where $x_1,x_2,\cdots,x_n$ are random samples of random variable $X$, n is the sample size, $K(\cdot)$ is the kernel smoothing function, and $h$ is the bandwidth [24]. In the current work, we use the *ksdensity* function in Statistics and Machine Learning Toolbox of Matlab to estimate the PDF functions, and the bandwidth parameter $h$ used is the default value which is proportional to $n^{-\frac{1}{5}}$. The estimation error of the KDE may be determined using methods reported by Silverman [24].

Boundary correction [25] is important for estimation of the PDF functions. Biomedical signals often have a heavy right tail distribution (e.g., power law distribution [26]) instead of normal distribution. In this case, log transformation is applied for data correction [27]. Data correction can also be modelled using logit transformation of $X$ as follows,

$$x_i \leftarrow \frac{1}{1+e^{-x_i}} \tag{5}$$

The values of the logit-transformed variable will be limited to the range [0,1], and distribution of the transformed variable will be closer to a normal distribution than the original variable [28]. The logit transformation step is recommended for the dataset with a small sample size or with outliers.

Data processing inequality (DPI) is an information theoretic concept which states that the information content of a signal cannot be increased via a local physical operation. It is an important property to describe the information flow on a network. DPI has been widely used in gene co-expression and cellular network analysis studies [29], [30]. DPI states that if the two sets $s_1$ and $s_3$ interact only through a third set $s_2$ (i.e. $s_1 \leftrightarrow s_2 \leftrightarrow s_3$), and if there is no alternative path between $s_1$ and $s_3$, then the mutual information $I\left(s_1, s_3\right), I\left(s_2,s_3\right)$ and $I\left(s_1,s_2\right)$ should satisfy the following inequality,

$$I(s_1,s_3) \leq \min\left[I\left(s_1,s_2\right); I\left(s_2,s_3\right)\right] \tag{6}$$

The set pair with the smallest MI in the triplet is likely to be indirectly connected in the network model. The connection between pair with lowest MI is eliminated among the triplet in DPI procedure, which is often used to enforce sparsity to a network model.

A standard way of enforcing sparsity to the association network is by using threshold pruning method. The method assumes that the associations of the edges around a node satisfy a similar distribution for all nodes on the whole network. In the current work, DPI is used instead of the threshold pruning method because the distribution of MI is heterogeneous on the network.

### B. DATA REDUCTION FOR METABOLITE SET ANALYSIS

Let $\boldsymbol{X} = (x_{ij})_{n\times m}$ be a concentration matrix with n samples and m metabolites from a metabolomics study. $\boldsymbol{X}$ is a metabolite level dataset that contains the concentrations of the detected metabolites. Assume that the detected metabolites are involved in B predefined metabolite sets. Typically, the metabolite sets are constructed based on metabolism or metabolic pathways in the KEGG or HumanCyc databases. For the rest of the article, we will use "pathway" and "metabolite set" interchangeably.

The metabolite level matrix $\boldsymbol{X}$ can be converted into a pathway level matrix $\boldsymbol{P}$ as follows. Assume that there are $L_b(0 < L_b \leq m)$ detected metabolites in the $b^{th}(b = 1, 2, \cdots, B)$ pathway. The submatrix $\boldsymbol{P^{(b)}} = (x_{ij})_{n\times L_b}$ can be obtained by concatenating the $L_b$ metabolites data (i.e., $L_b$ columns in $\boldsymbol{X}$) belonging to the $b^{th}$ pathway. Then the concatenated matrix $\boldsymbol{P} = (\boldsymbol{P^{(1)}}, \boldsymbol{P^{(2)}}, \cdots, \boldsymbol{P^{(B)}})$ is called the pathway-level matrix. Note that the pathway level matrix $\boldsymbol{P}$ has the same number of rows (samples) as $\boldsymbol{X}$, but may have a different number of columns (variables) from $\boldsymbol{X}$ as metabolites may be involved in several pathways.

The estimation of the PDF of multidimensional data requires a large number of samples. For an $L_b$-dimensional pathway dataset, a large number of samples are required to calculate the pathway-related entropy or mutual information. However, the number of samples in a metabolomics study is usually only a few hundred or often even less. For a dataset with small sample size, dimension reduction is an efficient and natural way to estimate the mutual information of two pathways. Principal components analysis (PCA) [31] is the most commonly used linear dimension reduction method in metabolomics. It maps high-dimensional data into a low-dimensional space through linear projection, and it is expected to retain as much data variation as possible in the projected dimension. In this work, we use PCA to project the multidimensional data from a pathway into a one-dimensional variable for subsequent calculations.

Let $\boldsymbol{P^{(b)}}$ be the submatrix (metabolite-based) of pathway $b$ with n samples and $L_b$ variables (metabolites), then its first principal component $t_{(b)}$ is given as follows,

$$\boldsymbol{P^{(b)}} = \boldsymbol{t}_{(\underline{\mathrm{b}})}\boldsymbol{l}^{\boldsymbol{T}}_{(b)} + \boldsymbol{R}, \quad b = 1, 2, \cdots, B \tag{7}$$

where $\boldsymbol{t}_{(b)}$ and $\boldsymbol{l}_{(b)}$ are the score vector and loading vector of the first principal component, respectively. $\boldsymbol{R}$ is the residuals of the PCA model of $\boldsymbol{P}^{(b)}$. After dimension reduction using PCA, the multi-dimensional pathway-level matrix is converted into a one-dimensional pathway-level dataset as follows,

$$\boldsymbol{T} = (\boldsymbol{t}_{(1)}, \boldsymbol{t}_{(2)}, \cdots, \boldsymbol{t}_{(B)}) \tag{8}$$

Then, MI of two pathways can be estimated by the dimension reduced pathway level dataset $\boldsymbol{T}$.

### C. METABOLITE SET ASSOCIATION NETWORK MODELLING

Let $\boldsymbol{P} = (\boldsymbol{P^{(1)}}, \boldsymbol{P^{(2)}}, \cdots, \boldsymbol{P^{(B)}})$ be a pathway level matrix with B metabolite sets (pathways) and n samples. The pathway level association network of the system can be represented by a graph $G\left(V, E\right)$ where node $v_i \in V$ represents pathway $\boldsymbol{P}^{(i)}$, and edge $e_{ij} \in E$ represents the association from pathway $\boldsymbol{P}^{(i)}$ to pathway $\boldsymbol{P}^{(j)}$. Here, we present a mutual information-based network modelling scheme to calculate edge $\boldsymbol{e}_{ij} \in E$ as follows

Step 1: Dimension Reduction. Perform dimension reduction on pathway-level data $\boldsymbol{P}$ to obtain a one-dimensional dataset $\boldsymbol{T} = (\boldsymbol{t}_{(1)}, \boldsymbol{t}_{(2)}, \cdots, \boldsymbol{t}_{(B)})$.

Step 2: MI Calculation. For each pathways pair $(i, j)$, calculate the probability density $\boldsymbol{p}(\boldsymbol{t}_{(i)})$ and $\boldsymbol{p}\left(\boldsymbol{t}_{(i)}, \boldsymbol{t}_{(j)}\right), (i = 1, 2, \cdots, B; j = 1, 2, \cdots, B)$ using the Gaussian kernel density estimator based on (4). Then, calculate the mutual information $\boldsymbol{I}_{ij}$ using (3) to obtain an MI matrix $\boldsymbol{I} = (I_{ij})_{B\times B}$.

Step 3: Link Pruning. Prune MI matrix $\boldsymbol{I}$ according to the DPI principle using (6) to obtain a sparse and symmetric matrix $\boldsymbol{E}$ as follows,

$$\boldsymbol{E} = \mathrm{DPI}(\boldsymbol{I}) \tag{9}$$

The pruned MI matrix $\boldsymbol{E}$ is the adjacent matrix of the network.

In the current study, the network modelling steps constitute the metabolite set association network (MSAN). In a MSAN network, nodes represent metabolite sets while edges represent the associations between two corresponding metabolite sets.

### D. DIFFERENTIAL NETWORK MODELLING

A differential pathway between an experimental group and a control group may be associated with the perturbation due to disease or treatment. Identification of the key differential pathways may provide new insight into the biological mechanisms related to disease or treatment. In this case, we have constructed a differential MSAN network (difMSAN) to model the difference between two biological states as follows.

Let $G^1$ and $G^2$ be two groups of samples, one is treated group with $n_1$ samples, another is control group with $n_2$ samples, and $\boldsymbol{X} = (x_{ij})_{(n_1+n_2)\times m}$ be the metabolite-level dataset of $G^1$ and $G^2$. After defining the metabolite sets as pathways predefined in a given database like KEGG, we can convert $\boldsymbol{X}$ into a pathway level matrix $\boldsymbol{P} = (\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \cdots, \boldsymbol{P}^{(B)})$, B is the number of the involved pathways (metabolite sets). Let $\boldsymbol{y} = (y_i)_{n\times 1}$, $n = n_1 + n_2$, be the response variable of the system, which is the class identity for the samples,

$$y_i = \begin{cases} 1 & \text{if sample } i \in G^1 \\ 0 & \text{if sample } i \in G^2 \end{cases}$$

Then the association between two metabolite sets (pathways) can also be quantified by the mutual information between these two pathways, which can be constructed by replacing the PCA-based dimension-reduced dataset $\boldsymbol{T}$ in the Step 1 of MSAN modelling procedure with partial least squares (PLS)-based dimension-reduced dataset as follows,

$$\boldsymbol{U} = (\boldsymbol{u}_{(1)}, \cdots, \boldsymbol{u}_{(b)}, \cdots, \boldsymbol{u}_{(B)}) \tag{10}$$

where $\boldsymbol{u}_{(b)}$ is the first latent variable of PLS model of pathway b in $G^1$ and $G^2$, which is calculated as follows [32],

$$\begin{aligned} \boldsymbol{P}^{(b)} &= \boldsymbol{t}_{(b)}\boldsymbol{l}_{(b)}^{\boldsymbol{T}} + \boldsymbol{R} \\ \boldsymbol{y} &= \boldsymbol{u}_{(b)}\boldsymbol{q}_{(b)}^{\boldsymbol{T}} + \boldsymbol{F} \end{aligned}$$

where $\boldsymbol{l}_{(b)}$ and $\boldsymbol{q}_{(b)}$ are the loading vectors, $\boldsymbol{t}_{(\boldsymbol{b})}$ and $\boldsymbol{u}_{(\boldsymbol{b})}$ are the score vectors, and $\boldsymbol{R}$ and $\boldsymbol{F}$ are the residual matrix of $\boldsymbol{P}^{(\boldsymbol{b})}$ and $\boldsymbol{y}$, respectively. More details refer to the literature [32].

In addition, we can calculate the mutual information between a given pathway $b$ and the response variable $\boldsymbol{y}$ as,

$$I_{(b)} = \sum p\left(\boldsymbol{u}_{(b)}, \boldsymbol{y}\right) \log\left(\frac{p\left(\boldsymbol{u}_{(b)}, y\right)}{p\left(\boldsymbol{u}_{(b)}\right) p\left(y\right)}\right) \tag{11}$$

The differential pathways between $G^1$ and $G^2$ can be identified taking both pathway and the edges it connects to into consideration in the difMSAN network.

## III. COLORECTAL CANCER METABOLOMICS DATASET

In this study, a published colorectal cancer (CRC) dataset [33] is used to evaluate the proposed method. MSAN networks are constructed and applied to reveal the key pathways which can be associated with colorectal cancer or the presence of polyps. The CRC dataset [33] has been deposited into the public repository of Metabolomics WorkBench (https://www.metabolomicsworkbench.org/) with Project ID: PR000226.

In brief, 234 volunteers underwent either colonoscopy or CRC surgery, and blood samples from the patients are obtained after overnight fasting and identical bowel preparation prior to the procedure. The volunteer samples are divided into three groups including healthy controls, CRC patients, and patients with colorectal polyps based on colonoscopy examination results. The serum samples collected included 66 CRC samples (Cancer), 76 polyp samples (Polyp), and 92 healthy control samples (Healthy). The experimental work is conducted in accordance with the protocols approved by the Indiana University School of Medicine and Purdue University Institutional Review Boards. A targeted LC−MS/MS approach is used for comprehensive CRC serum metabolic profiling under a standard operating procedure. In total, 113 metabolites out of 158 targeted MRM transitions are reliably detected, with a median QC coefficient of variance (CV) of 8% (see more details in [33]).

We define the metabolite sets as the metabolic pathways predefined in the public database of Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.kegg.jp/). In KEGG metabolic pathways database, human (Homo sapiens) metabolic pathways contain 1498 identified endogenous metabolites (excluding most lipids) and 85 pathways, each of which involves multiple metabolites. Detected metabolites that could not be matched with their corresponding metabolite ID in the KEGG dataset are removed from further analysis. The remaining 89 metabolites are associated with metabolic pathways based on the KEGG database. A total of 30 pathways with at least three detected metabolites are included for the analysis (TABLE 1). These 30 pathways consist of six categories, i.e., carbohydrate metabolism, lipid metabolism, amino acid metabolism, nucleotide metabolism, metabolism of other amino acids and metabolism of cofactors and vitamins.

Based on the KEGG databases, it should be noted that some of the pathways listed in TABLE 1 consists of a collection of metabolic pathways and not pathways themselves. This is not unusual in the current metabolic pathway analysis. It is mainly due to the common technological limitation in metabolomics studies with relatively low number of detected metabolites compared to total metabolome. Metabolite sets may also be defined based on metabolic pathways in the HumanCyc database (http://humancyc.org), however

**TABLE 1. List of 30 Pathways/metabolite set used in the colorectal cancer dataset.**

| Pathway Index | Pathway Name | Total Compd. | Detected Compd. |
|---|---|---|---|
| PI-1 | Glycolysis/Gluconeogenesis | 26 | 6 |
| PI-2 | Citrate cycle (TCA cycle) | 20 | 7 |
| PI-3 | Pentose phosphate pathway | 22 | 3 |
| PI-4 | Fructose and mannose metabolism | 20 | 5 |
| PI-5 | Galactose metabolism | 27 | 3 |
| PI-6 | Primary bile acid biosynthesis | 46 | 4 |
| PI-7 | Arginine biosynthesis | 14 | 7 |
| PI-8 | Purine metabolism | 65 | 11 |
| PI-9 | Pyrimidine metabolism | 39 | 6 |
| PI-10 | Alanine, aspartate & glutamate metabolism | 28 | 12 |
| PI-11 | Glycine, serine and threonine metabolism | 33 | 11 |
| PI-12 | Cysteine and methionine metabolism | 33 | 4 |
| PI-13 | Valine, leucine & isoleucine degradation | 39 | 6 |
| PI-14 | Valine, leucine & isoleucine biosynthesis | 8 | 4 |
| PI-15 | Lysine degradation | 25 | 3 |
| PI-16 | Arginine and proline metabolism | 38 | 8 |
| PI-17 | Histidine metabolism | 16 | 4 |
| PI-18 | Tyrosine metabolism | 42 | 7 |
| PI-19 | Tryptophan metabolism | 41 | 3 |
| PI-20 | Glutamine & glutamate metabolism | 6 | 3 |
| PI-21 | Glutathione metabolism | 28 | 4 |
| PI-22 | Starch and sucrose metabolism | 18 | 3 |
| PI-23 | Glycerolipid metabolism | 16 | 3 |
| PI-24 | Inositol phosphate metabolism | 30 | 3 |
| PI-25 | Glycerophospholipid metabolism | 36 | 4 |
| PI-26 | Pyruvate metabolism | 22 | 6 |
| PI-27 | Glyoxylate and dicarboxylate metabolism | 32 | 8 |
| PI-28 | Butanoate metabolism | 15 | 6 |
| PI-29 | Pantothenate and CoA biosynthesis | 19 | 3 |
| PI-30 | Porphyrin and chlorophyll metabolism | 30 | 3 |

**Note:** The total number of metabolites (Total Compd.) and number of detected metabolites (Detected Compd.) associated with each pathway are given in the last two columns

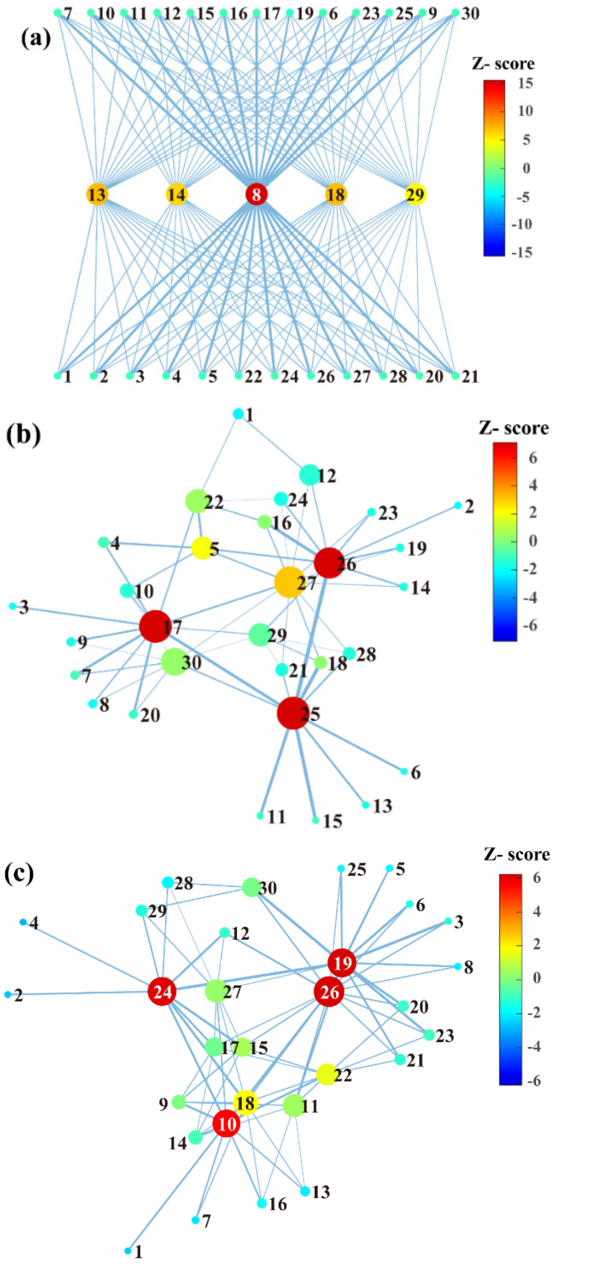


**FIGURE 1. MSAN networks of healthy (a), polyp (b) and cancer (c) groups. The pathway index is labeled in or near each circle. The linewidth of the edge represents the size of mutual information. The larger the mutual information, the larger the linewidth. The size of the circle correlates the betweenness centrality of the pathway in MSAN network. Color of the circle represents the z-score of the pathway with respect to random networks, which will be described in Section IV-B.**

the number of detected metabolites in pathways may be too low to warrant further investigation. If the metabolite set in the CRC study is defined based on the HumanCys database, only 63 pathway candidates and 71 detected metabolites (compared to 89 if the KEGG database is used) are left for the further analysis when we excluded the pathways with less than three metabolites (Supplementary information).

## IV. RESULTS AND DISCUSSION

### A. MSAN NETWORK CONSTRUCTION FOR THE CRC DATASET

Three MSAN networks of healthy, polyp and cancer groups are constructed as shown in Fig.1. The topology of the healthy MSAN network is found to be markedly different from the other two networks.

The five most important pathways which have the largest topological centrality (betweenness) in the healthy MSAN network are: *purine metabolism* (pathway index, PI-8), *valine, leucine and isoleucine degradation* (PI-13), *valine, leucine and isoleucine biosynthesis* (PI-14), *tyrosine metabolism* (PI-18), and *pantothenate and CoA biosynthesis* (PI-29). In the polyp MSAN network, *glycerophospholipid metabolism* (PI-25), *histidine metabolism* (PI-17), *pyruvate metabolism* (PI-26), and *Glyoxylate and dicarboxylate metabolism* (PI-27) are of larger betweenness centralities than the other pathways. The cancer MSAN network features in higher topological centralities of *pyruvate metabolism* (PI-26), *tryptophan metabolism* (PI-19), *alanine, aspartate and glutamate metabolism* (PI-10), and *inositol phosphate* *metabolism* (PI-24). These pathways have been previously identified as targets or reported to be perturbed by colon cancer [34], [35].

As the betweenness of these pathways is relatively large, they may play an important role in information flow in the network, which implies they may have a greater impact on the entire network when they are perturbed. For example, *pyruvate metabolism* (PI-26) is found in the polyp and cancer groups, which suggests that pyruvate metabolism acts as the core central pathway for MSAN network in abnormal states (polyps and cancer states) [34]. Interestingly, *tyrosine metabolism* (PI-18) is found to be a pivotal pathway in the healthy and cancer groups, but its betweenness in the polyp group prior to cancer development is much smaller. This suggested that when the body reaches an equilibrium state (non-transitional state), tyrosine metabolism may appear to act as a core pathway. Therefore, investigations focused on tyrosine metabolism may help to better understand the pathological process of certain cancers [36].

In addition, the structural stability of the MSAN network is evaluated using a bootstrapping method. The mutual information of each pathway pair is calculated using 80% of the samples bootstrapped from a given group. The coefficient of variance (CV) of each edge (pathway pair) is calculated with 200 bootstrapping procedures. The experimental result is presented in Figure S1 in the Supplemental Information. We found that the healthy MSAN network is of higher tolerability to the external interference than the other two networks, since most of the edges in the healthy MSAN networks are of large CV values. Similar results are reported in recent articles [35], in which smaller correlation coefficients are found between metabolites in healthy groups.

### B. TOPOLOGY ANALYSIS ON MSAN NETWORK

Then, the influence of each node on the structural topology of a MSAN network is analyzed and quantified using random walk with restart (RWR) process. RWR explores the global structure of the network by propagating the information along the network [37]. In the RWR process, the information diffuses away in the network from the initial distribution and the influence of each node can be quantified by the final or steady state of the network. The RWR algorithm is defined as reported in [37],

$$\boldsymbol{v}_t = \alpha \boldsymbol{v}_0 + (1-\alpha)\boldsymbol{W}\boldsymbol{v}_{t-1} \tag{12}$$

where $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{D}^{-1}$, $\boldsymbol{A}$ is the adjacency matrix, $\boldsymbol{D}$ is the diagonal degree matrix, and $\boldsymbol{v}_t$ is the state vector at step $t$. $\boldsymbol{v}_0$ is the initial state which is usually set to all ones if there is no prior information on the nodes, and $\alpha$ is the restart probability which is previously found to have only a slight effect on the results of the RWR [38], and $\alpha = 0.2$ is common setting in various studies. The iteration would reach a stationary state after a certain number of steps, and the final state $\boldsymbol{v}_t$ captures the global influence of nodes from the network, so $\boldsymbol{v}_t$ is the influence profile of the network. To simplify the notation, we will use "$\boldsymbol{v}_t$" and "$\boldsymbol{v}$" interchangeably in the rest of the article.

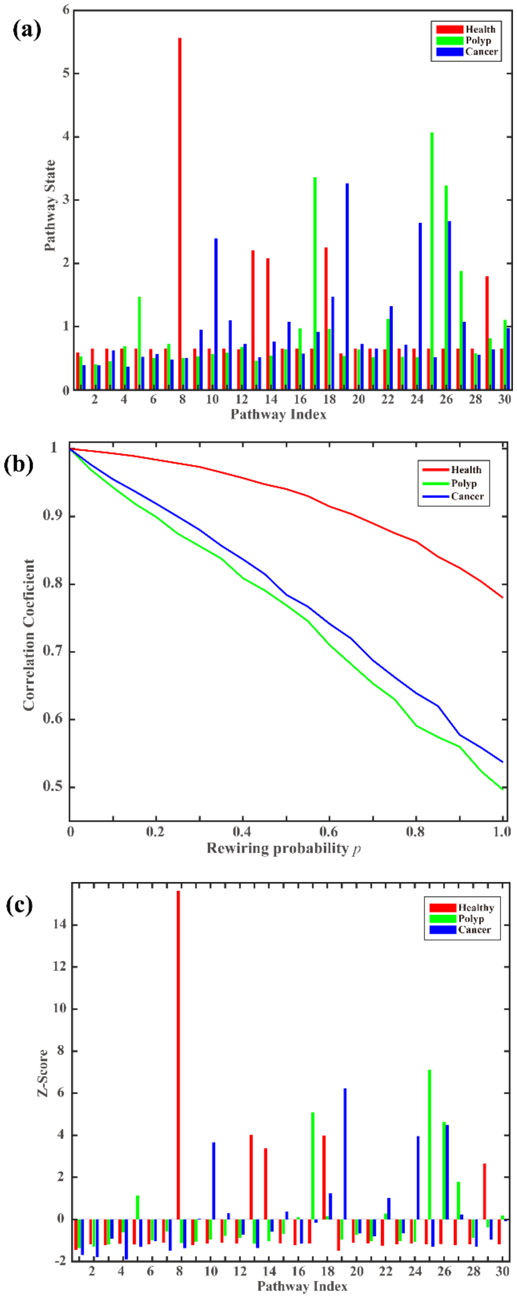


**FIGURE 2. Influence of each pathway after random walks on the MSAN network. (a) State distribution of pathways after RWR; (b) Correlation between the influence profiles of the networks before and after edge rewiring; (c) Z score of the influence profile of the pathways with respect to random networks.**

In our study, RWR process is carried out in the healthy, polyp and cancer MSAN networks, the influence profiles are obtained and shown in Fig.2a. The influence profiles are found consistent with the results topological centralities shown in Fig.1. Higher influence pathways are likely to be more topological importance in the network. For example, *purine metabolism* (PI-8) outperforms the other pathways in both the influence profile and the betweenness in the healthy MSAN network.

Structural stability, also called the anti-interference capability, is commonly used to evaluate the robustness or the sensitivity of the network when some connections (edges) are differential under certain circumstances, which is an important feature in topological analysis of a network. Here we evaluated both the structural stability of the overall network and the sensitivity of each pathway on the network.

Next, structural sensitivity of the overall network is evaluated by comparing the influence profile of the real network with that of the random networks (null models). Firstly, we generate a number of random networks by rewiring each edge on the MSAN network with different probabilities $p$, then performed RWR on the random networks to obtain the steady-states of the random networks. The initial distribution of the network is also set to $\boldsymbol{v}_0 = 1$ for all nodes. The correlation coefficients between the two steady-state distributions are calculated before and after edge rewiring. The experimental results are presented in Fig.2b, where the x-axis represents the probability of edge rewiring and y-axis represents the average correlation coefficients of 1000 random repeats. Fig.2b shows that the correlation coefficient between the steady-states (influence profile) of the MSAN network and that of the random network decreases with an increasing of edge rewiring probability.

As expected, the healthy network showed the strongest structural stability, while the polyp network is characterized by the weakest structural stability. When 100% edges are reconnected, a Pearson correlation coefficient of up to 0.8 is observed for the healthy MSAN network, but it is only 0.5 for the polyp MSAN network. Two possible reasons may explain these different structural stability results. First, the healthy MSAN network has the highest number of edges, while the polyp MSAN network has the least (124, 72 and 56 edges for healthy, cancer and polyp MSAN networks, respectively). It is plausible that a larger number of edges led to a stronger structural stability. Second, from a systems dynamic point of view, the healthy condition can be regarded as an equilibrium state with a large steady-state space, while cancer might represent another equilibrium state. On the other hand, the polyp group may represent a less stable transition state between healthy and cancer states. Therefore, the metabolic network of healthy people may have better anti-interference capability than that for patients.

The sensitivity of a pathway in a MSAN network is evaluated by comparing its topological influence with that of in the random networks. A total of 1000 random networks are generated by rewiring each edge on the MSAN network with a rewiring probability $p = 0.5$, then RWR is performed on the edge rewired networks to obtain the steady-state from an initial state $v_0 = 1$ for all nodes. Z-score of the steady-state of each node of the given MSAN network is calculated with respect to that of the edge rewired networks as follows,

$$Z_b = (v_b^{\text{real}} - \bar{v}_b^{\text{rand}})/\sigma\left(v_b^{\text{rand}}\right) \tag{13}$$

where $v_b^{real}$ is steady-state of pathway $b$ on the real MSAN network, $\bar{v}_b^{\text{rand}}$ and $\sigma\left(v_b^{\text{rand}}\right)$ are the mean and standard deviation of the steady-state of pathway $b$ on the simulated random networks, respectively. Pathways with positive z-score values are considered over-represented in the network, while pathways with negative Z-scores are under-represented in the network. Pathway with $|\text{Z-score}| > 2$ is considered significantly sensitive to the interference. (Z-score of the nodes are shown using a color scale in Fig.1).

As shown in Fig.2c, the pathways with higher topological centralities in the real MSAN network are found more likely to be over-represented. For example, the cancer MSAN network over-represented four pathways including *tryptophan metabolism* (PI-19), *pyruvate metabolism* (PI-26), *inositol phosphate metabolism* (PI-24), and *alanine, aspartate & glutamate metabolism* (PI-10). On the other hand, the polyp MSAN network only over-represented three pathways including *glycerophospholipid metabolism* (PI-25), *histidine metabolism* (PI-17), and *pyruvate metabolism* (PI-26). While in the healthy MSAN network, the five pathways with high topological centralities are all found over-represented. There is no pathway found to be under-represented in the three MSAN networks.

## C. COMPARISON OF DIFFERENT METHODS ON IDENTIFICATION OF KEY PATHWAYS

### 1) KEY PATHWAYS IDENTIFIED BY DIFFERENTIAL MSAN NETWORK

Three difMSAN networks of polyp vs. healthy, cancer vs. polyp, and cancer vs. healthy are constructed and visualized in Fig.3. Permutation test is used to quantify the significance of the node and edge in a given difMSAN network as follows:

Step 1. Randomly permute the response variable $\boldsymbol{y}$

Step 2. Calculate mutual information of each node and each edge using the permuted $y$, then build a random network using the MI matrix of the edges;

Step 3. Run RWR on the random network by setting the initial state vector $\boldsymbol{v}_0 = (I_b)_{B\times 1}$ and the restart probability $\alpha = 0.5$, where $I_b$ is the mutual information of node b;

Step 4. Repeat Step $1 \sim 3$ for 200 times to generate 200 random difMSAN networks and their steady-state $v_t$;

Step 5. Calculate Z-score for each edge by,

$$Z_i = (I_i^{\text{real}} - \bar{I}_i^{\text{rand}})/\sigma\left(I_i^{\text{rand}}\right) \tag{14}$$

where $I_i^{\text{real}}$ is the MI value of edge $i$ in the real difMSAN network, and $\bar{I}_i^{\text{rand}}$ and $\sigma\left(I_i^{\text{rand}}\right)$ are the mean and standard deviation of the MI values in the random difMSAN networks, respectively; Then, calculate Z-score of $\boldsymbol{v}_t$ for each node according to (13);

Step 6. Calculate the computational p-value for each node and each edge in the difMSAN network using their Z-score values by assuming that the distribution of MI of the edge or $\boldsymbol{v}_t$ of the node be normal in the random networks (null models).

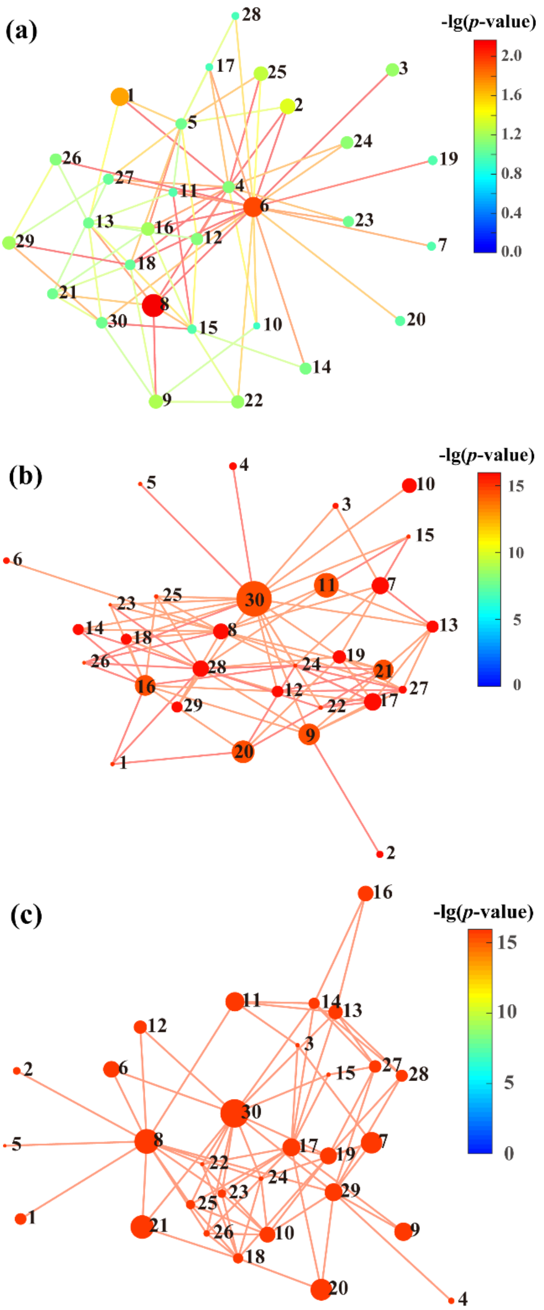


**FIGURE 3.** MSAN networks of (a) Polyp versus healthy; (b) Cancer versus polyp; (c) Cancer versus healthy. The pathway index is labeled in or near each circle. The size of the circle correlates the mutual information between the pathway and the response variable *y*. Color of the circle and the edge represents the significance of difference.

The significance of each node and each edge is highlighted on the differential MSAN network, as shown in Fig.3. There are only a few significant difference (edges or nodes) between polyp and healthy groups (Fig.3a). the results suggested limited perturbation in polyp state. The first two highest z-score pathways are found to be *purine metabolism* (PI-8, $p$-value $= 0.03$) and *primary bile acid biosynthesis* (PI-6, $p$-value $= 0.09$). Six edges are of significant difference ($p$-value $< 0.05$), and the most significant edge ($p$-value $=$ $0.007$) is the edge between *pyruvate metabolism* (PI-26) and *primary bile acid biosynthesis* (PI-6). The finding suggests interaction between the two pathways may be altered in polyp patients.

Most of the edges and nodes are found to be significantly different in both the differential networks of cancer vs. polyp (Fig.3b) and cancer vs. healthy (Fig.3c). The results indicate metabolism of cancer state deviated greatly from the polyp or healthy states. Fig.3c shows that five pathways including *porphyrin and chlorophyll metabolism* (PI-30), *purine metabolism* (PI-8), *glutathione metabolism* (PI-21), *glutamine & glutamate metabolism* (PI-20), and *arginine biosynthesis* (PI-7) exhibited greater influence in the differential MSAN network comparing cancer versus healthy. It implied that these pathways are differential for cancer group with respect to the healthy group. These results are further supported by a number of recent studies [39]–[47]. For example, Sheng and coworkers [39] found the *pathway of porphyrin and chlorophyll metabolism* to be significantly altered in CRC patients, the perturbation is also validated in literature [40] by the genes from the pathway. Purines are basic components of nucleotides in cell proliferation, thus impaired *purine metabolism* has been associated with the progression of cancer [41], [42]. *Glutathione* (GSH) is the most abundant antioxidant found in living organisms and has multiple functions, most of which maintain cellular redox homeostasis [43]. Bansal and Simon [44] reviewed the recent studies on deciphering the role of GSH in tumor initiation and progression as well as mechanisms underlying how GSH imparts treatment resistance to growing cancers. *Glutamine and glutamate metabolism* plays key roles in tumor growth and invasion, and this is also reported in numerous CRC studies [45], [46]. *Arginine* is critical for the growth of human cancers; it is involved in diverse aspects of tumor metabolism. Delage and coworkers [47] showed arginine deprivation and argininosuccinate synthetase expression in the treatment of colorectal cancer.

### 2) KEY PATHWAYS IDENTIFIED BY MetaboAnalyst

MetaboAnalyst 4.0 (http://www.metaboanalyst.ca) is a publicly accessible software platform commonly used in metabolomics studies. MetaboAnalyst 4.0 provides a pathway analysis module [48], in which the underlying network is the real-world reaction network from the KEGG dataset, *i.e.*, a metabolite level network. MetaboAnalyst 4.0 evaluates the impact of a pathway using its topological centrality on the network, such as betweenness, closeness and degree. A larger pathway impact signifies a more important pathway is. It also provides some metrics based on ORA, e.g., fold-enrichment and Fisher's exact test. Here we identify the key pathways for the three groups using MetaboAnalyst 4.0, as shown in Fig.4.

As shown in Fig.4a, the top five relevant pathways between the healthy and cancer groups are: arginine biosynthesis; alanine, aspartate and glutamate metabolism; D-glutamine and D-glutamate metabolism; arginine and proline metabolism; and histidine metabolism. The findings are different from the results of the differential MSAN network analysis. There are

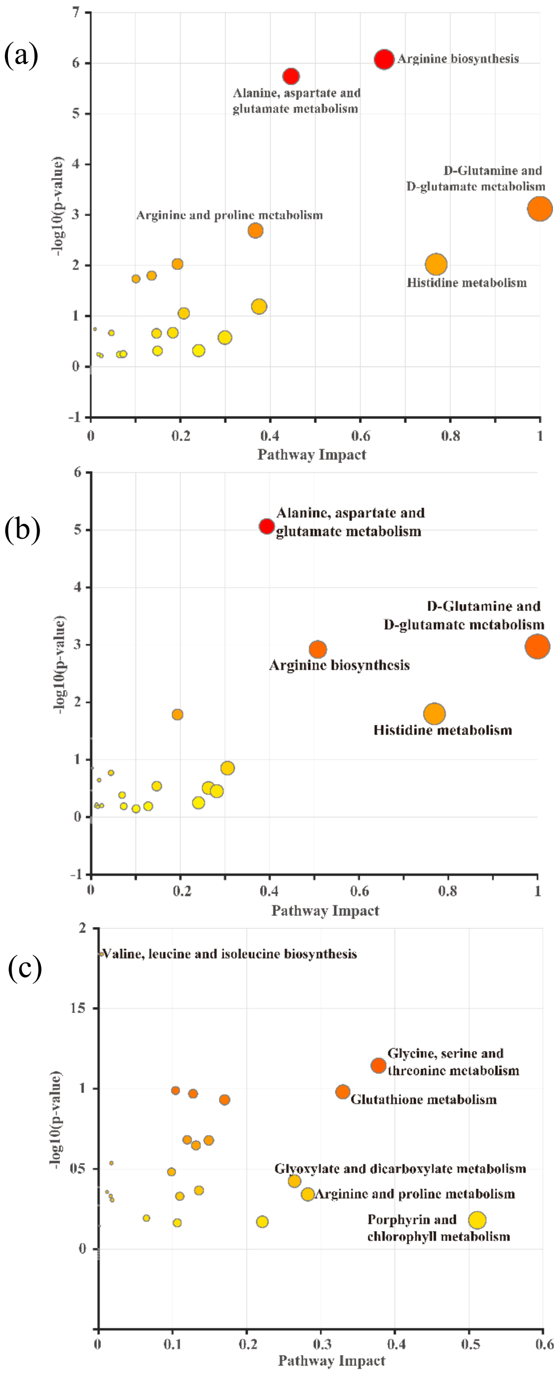


**FIGURE 4.** **Analysis comparing two study groups. (a) Cancer versus healthy; (b) Cancer versus polyp; (c) Polyp versus healthy. The pathway impact along the abscissa is calculated by topological analysis with the relative betweenness centrality; p-values are the result of Fisher's Exact test.**

two main possible reasons for these differences. On the one hand, ORA screens significantly changed metabolites before calculating pathway impact and Fisher's exact test, and the screening criteria affect the subsequent results. For example, the results of ORA varied with different values of the VIP parameter, VIP > 1.0 or VIP > 1.5. As the selection of these thresholds is mostly empirical and there is no uniform standard, it might impact the reliability of the ORA results. On the other hand, the number of metabolites detected is also a major determinant of the ORA results. This is because the calculation of pathway impact and Fisher's Exact test is based on the hit number of significant metabolites in a pathway. Thus, ORA is more likely to select pathways with high hit ratios while underestimate pathways with low hit ratio but high significance ratio. For example, tryptophan metabolism varies greatly in healthy and cancer groups in the results of differential MSAN network, but ORA did not highlight the pathway as there are only three out of 41 possible metabolites detected in this pathway. *Inositol phosphate metabolism*, which has 30 metabolites in total but only 3 detected metabolites in the current dataset, is found to be a significant pathway based on MSAN network analysis but also appeared to be not significant in the ORA results. Most of the highly significant pathways in ORA contain a high rate of detected metabolites. For example, in the first three pathways, the metabolite detection rate is 50% (7 of 14) for arginine biosynthesis, 42.86% (12 of 28) for alanine, aspartate and glutamate metabolism, and 50% (3 of 6) for Glutamine and Glutamate metabolism. However, the number of detected metabolites depends mainly on the sensitivity of the analytical instrument and the experimental parameter settings, and is independent of the disease state. The ORA algorithm relies on the number of detected metabolites, which may lead to bias in its selection of important pathways. We believe the results of the proposed MSAN method may complement the conventional ORA results, and the results from both analysis may converge with increasing number of detected metabolites in a pathway.

### 3) KEY PATHWAYS IDENTIFIED BY NetGSA

Another commonly used pathway analysis tool, NetGSA [49], incorporates network relationships into pathway analysis. When the network information is incomplete, a probabilistic graphical model is used to complete the pathway topology based on the available data, while using the existing topology information as constraints [9]. Here, we first construct a Gaussian graphical model (GGM) [50] of the metabolite-level network following the instruction of the R-package of NetGSA, then conduct NetGSA method based on the GGM model. We use samples in each comparison group to construct network, i.e. healthy samples were used to construct the healthy network and cancer samples were used to construct the cancer network. We have not used pooled samples for a single network construction due to heterogeneity in the data which may cause bias on the result. Fig.5 shows the network obtained from GGM, with the top five important pathways listed in TABLE 2.

As shown in TABLE 2, the top five selected pathways between the healthy and cancer groups include: Butanoate biosynthesis; Arginine biosynthesis; Purine metabolism; Citrate cycle (TCA cycle); and Pyruvate metabolism, which are different from the previous results. Compared with ORA, NetGSA incorporates the network structure of metabolites and the information of network structure is constructed using

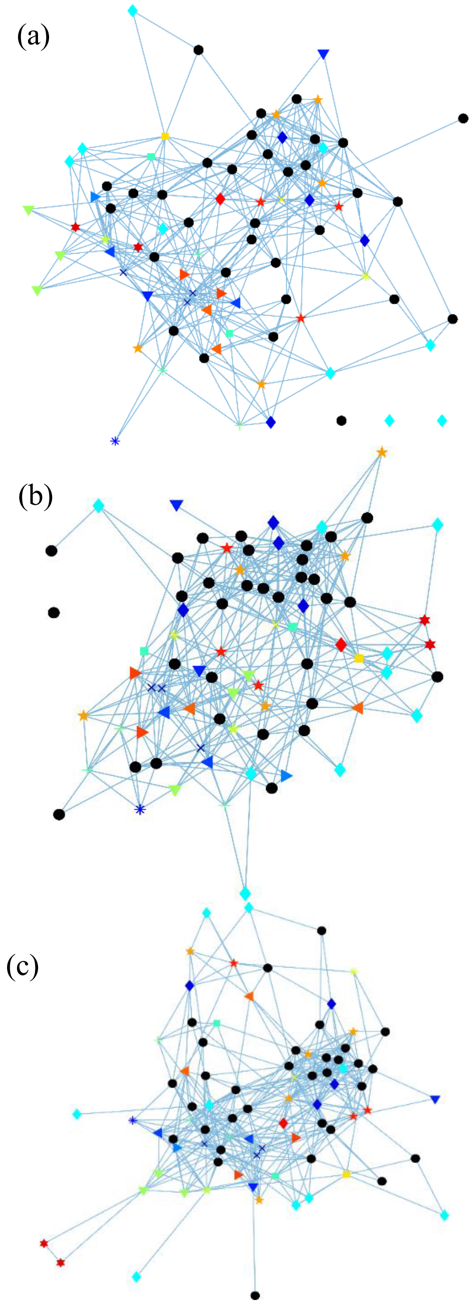


**FIGURE 5.** Metabolite level network by the Gaussian graph model approach. (a) Cancer versus healthy; (b) Cancer versus polyp; (c) Polyp versus healthy. Points in black are metabolites which belong to more than one pathway. For the other points, those with same color and same marker are the metabolites from a same pathway.

a propagated effect on each other metabolite through the influence matrix. This approach leads to more biologically meaningful results. However, NetGSA relies on the given network structure and the parameters in GGM. It is common that the metabolites in a given pathway cannot be fully detected in a metabolomics analysis. Thus, it is a challenging issue to construct a constrained network and estimation of a robust and meaningful GGM network in NetGSA. In our proposed method, instead of limiting the analysis to the metabolite level, we provide further analysis at the pathway level. The pathway interaction is constructed using estimation of mutual information and DPI, which not only took the non-linear effect into account but also alleviated the background network information.

**TABLE 2.** Top five important pathways identified by NetGSA.

| Rank | Cancer vs. Healthy | Cancer vs. Polyp | Polyp vs. Healthy |
|---|---|---|---|
| 1 | Butanoate metabolism | Valine, leucine and isoleucine degradation | Arginine and proline metabolism |
| 2 | Arginine biosynthesis | Arginine and proline metabolism | Glycine, serine and threonine metabolism |
| 3 | Purine metabolism | Purine metabolism | Arginine biosynthesis |
| 4 | Citrate cycle (TCA cycle) | Tyrosine metabolism | Tyrosine metabolism |
| 5 | Pyruvate metabolism | Butanoate metabolism | Valine, leucine and isoleucine degradation |

Note: The pathways that appear in top five ranks for more than one comparison are shaded with same colors.

## V. CONCLUSION

Network-based pathway analysis represents a new generation of pathway analysis methods which involves network construction and topology analysis. To date, the number of detected metabolites is still far lower than the actual number due to limitations inherent in the analytical platforms used in metabolomics. Thus, it remains a challenge to identify and interpret the differential pathways based on the limited number of metabolites. The results of the identified differential pathways may also differ based on the selected methods. In addition, the network-based pathway analysis methods currently in use are mainly based on metabolite-level networks, and the connections and interactions between pathways are often not taken into consideration.

In the current study, we develop the MSAN-based analysis method and apply to a published CRC metabolomics dataset. The method constructed a MSAN network where the pathway is represented as a node, and the irreducible statistical dependency between two pathways as an edge. From the perspective of biological information, the MSAN network may provide useful insights into the linkages between pathways. In addition, disease-related pathways can be evaluated by differential MSAN network analysis between disease patients and the controls. The results from the analysis of a CRC metabolomics dataset suggest that MSAN may reflect important biological states of the samples set. MSAN may help identify pathways that have a greater impact on the network and provide insight into changes in pathways across different disease states.

## ACKNOWLEDGMENT

*(Yueyue Liu and Xiangnan Xu contributed equally to this work.)*

## REFERENCES

[1] J. Ma, A. Shojaie, and G. Michailidis, "Network-based pathway enrichment analysis with incomplete network information," *Bioinformatics*, vol. 32, no. 20, pp. 3165–3174, Oct. 2016.

[2] N. Städler and S. Mukherjee, "Multivariate gene-set testing based on graphical models," *Biostatistics*, vol. 16, no. 1, pp. 47–59, Jan. 2015.

[3] W. N. van Wieringen, C. F. W. Peeters, R. X. de Menezes, and M. A. van de Wiel, "Testing for pathway (in) activation by using Gaussian graphical models," *Appl. Stat.-J. Roy. Stat. Soc.*, vol. 67, no. 5, pp. 1419–1436, Nov. 2018.

[4] Z. Gu, J. Liu, K. Cao, J. Zhang, and J. Wang, "Centrality-based pathway enrichment: A systematic approach for finding significant pathways dominated by key genes," *BMC Syst. Biol.*, vol. 6, no. 1, Jun. 2012, Art. no. 56.

[5] M. A.-H. Ibrahim, S. Jassim, M. A. Cawthorne, and K. Langlands, "A topology-based score for pathway enrichment," *J. Comput. Biol.*, vol. 19, no. 5, pp. 563–573, May 2012.

[6] L. Jacob, P. Neuvial, and S. Dudoit, "More power via graph-structured tests for differential expression of gene networks," *Ann. Appl. Statist.*, vol. 6, no. 2, pp. 561–600, Jun. 2012.

[7] M. S. Massa, M. Chiogna, and C. Romualdi, "Gene set analysis exploiting the topology of a pathway," *BMC Syst. Biol.*, vol. 4, no. 1, p. 121, Sep. 2010, doi: 10.1186/1752-0509-4-121.

[8] D. Wu and G. K. Smyth, "Camera: A competitive gene set test accounting for inter-gene correlation," *Nucleic Acids Res.*, vol. 40, no. 17, p. e133, Sep. 2012.

[9] J. Ma, A. Shojaie, and G. Michailidis, "A comparative study of topology-based pathway enrichment analysis methods," *BMC Bioinf.*, vol. 20, no. 1, Nov. 2019, Art. no. e546.

[10] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: Current approaches and outstanding challenges," *PLoS Comput. Biol.*, vol. 8, no. 2, Feb. 2012, Art. no. e1002375.

[11] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[12] H. S. Leong and D. Kipling, "Text-based over-representation analysis of microarray gene lists with annotation bias," *Nucleic Acids Res.*, vol. 37, no. 11, p. e79, Jun. 2009.

[13] J. J. Goeman and P. Buhlmann, "Analyzing gene expression data in terms of gene sets: Methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, Apr. 2007.

[14] C. Webber, "Functional enrichment analysis with structural variants: Pitfalls and strategies," *Cytogenetic Genome Res.*, vol. 135, nos. 3–4, pp. 277–285, 2011.

[15] J. Reimand, R. Isserlin, V. Voisin, M. Kucera, C. Tannus-Lopes, A. Rostamianfar, L. Wadi, M. Meyer, J. Wong, C. Xu, D. Merico, and G. D. Bader, "Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, cytoscape and EnrichmentMap," *Nature Protocols*, vol. 14, no. 2, pp. 482–517, Feb. 2019.

[16] I. Ihnatova and E. Budinska, "ToPASeq: An R package for topology-based pathway analysis of microarray and RNA-seq data," *BMC Bioinf.*, vol. 16, no. 1, Oct. 2015, Art. no. 350.

[17] X. Dong, Y. Hao, X. Wang, and W. Tian, "LEGO: A novel method for gene set over-representation analysis by incorporating network-based gene weights," *Sci. Rep.*, vol. 6, no. 1, Jan. 2016, Art. no. 18871.

[18] Y. Xu, X. Y. Yi, and H. M. Yue, "Dysregulated pathways identification analysis in Parkinson disease based on attractor of within-pathway effects and crosstalk inter-pathways," *Int. J. Clin. Express Med.*, vol. 10, no. 2, pp. 3079–3087, 2017.

[19] J. Han, C. Li, H. Yang, Y. Xu, C. Zhang, J. Ma, X. Shi, W. Liu, D. Shang, Q. Yao, Y. Zhang, F. Su, L. Feng, and X. Li, "A novel dysregulated pathway-identification analysis based on global influence of within-pathway effects and crosstalk between pathways," *J. Roy. Soc. Interface*, vol. 12, no. 102, Jan. 2015, Art. no. 20140937.

[20] M. L. Slattery, L. E. Mullany, L. Sakoda, W. S. Samowitz, R. K. Wolff, J. R. Stevens, and J. S. Herrick, "The NF-$\kappa$B signalling pathway in colorectal cancer: Associations between dysregulated gene and miRNA expression," *J. Cancer Res. Clin. Oncol.*, vol. 144, no. 2, pp. 269–283, Feb. 2018.

[21] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol.*, vol. 5, no. 1, pp. 54–66, Jan. 2007.

[22] K. A. Zielińska and V. L. Katanaev, "Information theory: New look at oncogenic signaling pathways," *Trends Cell Biol.*, vol. 29, no. 11, pp. 862–875, Nov. 2019.

[23] S. Calonico, M. D. Cattaneo, and M. H. Farrell, "Nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference," *J. Stat. Softw.*, vol. 91, no. 8, pp. 1–35, 2019, doi: 10.18637/jss.v091.i08.

[24] B. W. Silverman, *Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability*. London, U.K.: Chapman & Hall, 1986.

[25] T. Berry and T. Sauer, "Density estimation on manifolds with boundary," *Comput. Statist. Data Anal.*, vol. 107, pp. 1–17, Mar. 2017.

[26] S. Sato, M. Horikawa, T. Kondo, T. Sato, and M. Setou, "A power law distribution of metabolite abundance levels in mice regardless of the time and spatial scale of analysis," *Sci. Rep.*, vol. 8, no. 1, Jul. 2018, Art. no. e10315.

[27] J. P. Ekwaru and P. J. Veugelers, "The overlooked importance of constants added in log transformation of independent variables with zero values: A proposed approach for determining an optimal constant," *Statist. Biopharmaceutical Res.*, vol. 10, no. 1, pp. 26–29, Jan. 2018.

[28] D. Curran-Everett, "Explorations in statistics: The log transformation," *Adv. Physiol. Edu.*, vol. 42, no. 2, pp. 343–347, Jun. 2018.

[29] E. Eskandari, F. Mahjoubi, and J. Motalebzadeh, "An integrated study on TFs and miRNAs in colorectal cancer metastasis and evaluation of three co-regulated candidate genes as prognostic markers," *Gene*, vol. 679, pp. 150–159, Dec. 2018.

[30] R. W. Li, S. Wu, C.-J. Li, W. Li, and S. G. Schroeder, "Splice variants and regulatory networks associated with host resistance to the intestinal worm cooperia oncophora in cattle," *Veterinary Parasitol.*, vol. 211, nos. 3–4, pp. 241–250, Jul. 2015.

[31] Y. Shiokawa, Y. Date, and J. Kikuchi, "Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet," *Sci. Rep.*, vol. 8, no. 1, Feb. 2018, Art. no. e3426.

[32] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometric Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, Oct. 2001.

[33] J. Zhu, D. Djukovic, L. Deng, H. Gu, F. Himmati, E. G. Chiorean, and D. Raftery, "Colorectal cancer detection using targeted serum metabolic profiling," *J. Proteome Res.*, vol. 13, no. 9, pp. 4120–4130, Sep. 2014.

[34] J. Gu, Y. Xiao, D. Shu, X. Liang, X. Hu, Y. Xie, D. Lin, and H. Li, "Metabolomics analysis in serum from patients with colorectal polyp and colorectal cancer by 1H-NMR spectrometry," *Disease Markers*, vol. 2019, Apr. 2019, Art. no. 3491852, doi: 10.1155/2019/3491852.

[35] M. Afzal, E. Saccenti, M. B. Madsen, M. B. Hansen, O. Hyldegaard, S. Skrede, V. A. P. M. dos Santos, A. Norrby-Teglund, and M. Svensson, "Integrated univariate, multivariate, and correlation-based network analyses reveal metabolite-specific effects on bacterial growth and biofilm formation in necrotizing soft tissue infections," *J. Proteome Res.*, vol. 19, no. 2, pp. 688–698, Feb. 2020.

[36] M. Tong, T. L. Wong, S. T. C. Luk, N. Che, X. Y. Guan, Y. F. Yuan, T. K. W. Lee, and S. Ma, "Deranged tyrosine metabolism drives tumorigenesis in liver cancer," *Cancer Res.*, vol. 78, no. 13, Jul. 2018.

[37] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: A universal amplifier of genetic associations," *Nature Rev. Genet.*, vol. 18, no. 9, pp. 551–562, Sep. 2017.

[38] J.-Q. Cao, C.-X. Li, R.-Y. Wang, J.-J. Chen, S.-M. Ma, W.-Y. Wang, and L.-J. Meng, "Identification of atherosclerosis-related prioritizing metabolites based on a multi-omics composite network," *Experim. Therapeutic Med.*, vol. 17, no. 5, pp. 3391–3398, Mar. 2019.

[39] Q.-S. Sheng, K.-X. He, J.-J. Li, Z.-F. Zhong, F.-X. Wang, L.-L. Pan, and J.-J. Lin, "Comparison of gut microbiome in human colorectal cancer in paired tumor and adjacent normal tissues," *OncoTargets Therapy*, vol. 13, pp. 635–646, Jan. 2020.

[40] P. Ravindranathan, D. Pasham, U. Balaji, J. Cardenas, J. Gu, S. Toden, and A. Goel, "A combination of curcumin and oligomeric proanthocyanidins offer superior anti-tumorigenic properties in colorectal cancer," *Sci. Rep.*, vol. 8, no. 1, Sep. 2018, Art. no. e13869.

[41] J. Yin, W. Ren, X. Huang, J. Deng, T. Li, and Y. Yin, "Potential mechanisms connecting purine metabolism and cancer therapy," *Frontiers Immunol.*, vol. 9, Jul. 2018, Art. no. 1697.

[42] Y. Long, B. Sanchez-Espiridion, M. Lin, L. White, L. Mishra, G. S. Raju, S. Kopetz, C. Eng, M. A. T. Hildebrandt, D. W. Chang, Y. Ye, D. Liang, and X. Wu, "Global and targeted serum metabolic profiling of colorectal cancer progression," *Cancer*, vol. 123, no. 20, pp. 4066–4074, Jun. 2017.

[43] A. D. Kim, R. Zhang, X. Han, K. A. Kang, M. J. Piao, Y. H. Maeng, W. Y. Chang, and J. W. Hyun, ''Involvement of glutathione and glutathione metabolizing enzymes in human colorectal cancer cell lines and tissues,'' *Mol. Med. Rep.*, vol. 12, no. 3, pp. 4314–4319, Sep. 2015.

[44] A. Bansal and M. C. Simon, ''Glutathione metabolism in cancer progression and treatment resistance,'' *J. Cell Biol.*, vol. 217, no. 7, pp. 2291–2298, Jul. 2018.

[45] Z. Song, B. Wei, C. R. Lu, P. Y. Li, and L. Chen, ''Glutaminase sustains cell survival via the regulation of glycolysis and glutaminolysis in colorectal cancer,'' *Oncol. Lett.*, vol. 14, no. 3, pp. 3123–3317, Sep. 2017.

[46] G. Liu, J. Zhu, M. Yu, C. Cai, Y. Zhou, M. Yu, Z. Fu, Y. Gong, B. Yang, Y. Li, Q. Zhou, Q. Lin, H. Ye, L. Ye, X. Zhao, Z. Li, R. Hen, F. Han, C. Tang, and B. Zeng, ''Glutamate dehydrogenase is a novel prognostic marker and predicts metastases in colorectal cancer patients,'' *J. Translational Med.*, vol. 13, no. 1, May 2015, Art. no. 144.

[47] B. Delage, D. A. Fennell, L. Nicholson, I. McNeish, N. R. Lemoine, T. Crook, and P. W. Szlosarek, ''Arginine deprivation and argininosuccinate synthetase expression in the treatment of cancer,'' *Int. J. Cancer*, vol. 126, no. 12, pp. 2762–2772, Jun. 2010.

[48] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D. S. Wishart, and J. Xia, ''MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis,'' *Nucleic Acids Res.*, vol. 46, no. W1, pp. W486–W494, Jul. 2018.

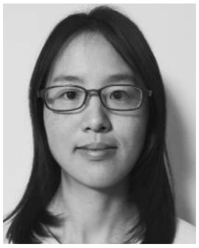
[49] A. Shojaie and G. Michailidis, ''Analysis of gene sets based on the underlying regulatory network,'' *J. Comput. Biol.*, vol. 16, no. 3, pp. 407–426, Mar. 2009.

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

**YUEYUE LIU** received the B.S. degree in electronic information science and technology from Xiamen University, China, in 2017, where she is currently pursuing the M.S. degree in physical electronics. Her current research interests include network analysis and machine learning and their applications in metabolomics.

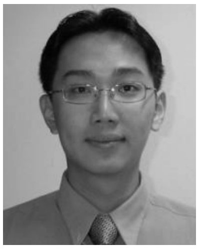
**XIANGNAN XU** was born in Pingdingshan, Henan, China, in 1992. He received the B.S. degree in information engineering from the Nanjing University of Aeronautics and Astronautics, in 2014, and the M.S. degree in electronic and communication engineering from Xiamen University, China, in 2017. He is currently pursuing the Ph.D. degree in bioinformatics with The University of Sydney, Australia.

His research interests include the analysis of high-throughput biotechnological data including that from NMR, MS, and next-generation sequencing and centered on the development of methods and the application of statistics to problems in-omics and biomedical research.



**LINGLI DENG** received the B.S. degree in communication engineering from Dalian Maritime University, Liaoning, China, in 2009, and the Ph.D. degree in communication and information systems from Xiamen University, Fujian, China, in 2015.

She is currently an Assistant Professor with the School of Information Engineering, East China University of Technology, Jiangxi, China. Her research interests include bioinformatics and machine learning.

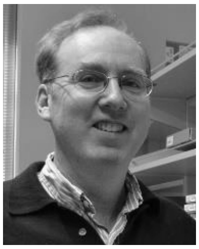


**KIAN-KAI CHENG** received the B.S. and M.S. degrees in bioprocess engineering from Universiti Teknologi Malaysia, in 2002 and 2005, respectively, and the Ph.D. degree in biochemistry from the University of Cambridge, in 2010.

He is currently a Senior Lecturer with Universiti Teknologi Malaysia. Since 2019, he has been the Director of the Innovation Centre in Agritechnology, Universiti Teknologi Malaysia Pagoh Research Centre. His research interests include metabolomics, systems biology, and multivariate data analysis.

**JINGJING XU** received the B.S. degree in electronic information science and technology and the Ph.D. degree in radio physics from Xiamen University, Xiamen, China, in 2005 and 2011, respectively.

From 2011 to 2014, she was an Engineer with the Department of Electronic Science, Xiamen University, where she has been a Senior Engineer, since 2015. Her research interests include the applications of metabolomics on therapy mechanisms of acupuncture and moxibustion and development of preprocessing methods on NMR-based metabolomics.



**DANIEL RAFTERY** received the A.B. degree from Harvard College and the Ph.D. degree in physical chemistry from UC Berkeley, in 1991.

From 1992 to 1994, he was a NSF Postdoctoral Scientist with the University of Pennsylvania. Then, he became a Faculty Member with the Department of Chemistry, Purdue University. In 2012, he moved to the University of Washington, where he is currently a Medical Education and Research Professor with the Department of Anesthesiology and Pain Medicine. He was a Purdue University Faculty Scholar, in 2006. He is the author of over 200 research articles and book chapters, two edited books, and ten patents. His research interests include the development and application of new methods in metabolomics and the discovery and validation of disease metabolite biomarkers. He was a recipient of the Research Corporation Cottrell Scholars Award, in 1997, the NSF Career Award, in 1998, the Alfred P. Sloan Foundation Award, in 1999. He is the Director of the Northwest Metabolomics Research Center and an Associate Editor of *Analytical Chemistry*.

**JIYANG DONG** was born in Quanzhou, Fujian, China, in 1974. He received the B.S. and M.S. degrees in physics education and optical engineering from Fujian Normal University, China, in 1994 and 1998, respectively, and the Ph.D. degree in condensed matter physics from Xiamen University, China, in 2001.

From 2001 to 2004, he was a Postdoctoral Scientist and an Associate Professor with the National Key Laboratory for Radar Signal Processing, Xidian University, China. From 2004 to 2011, he was an Associate Professor with the Department of Physics, Xiamen University. Since 2011, he has been a Professor with the Department of Electronic Science, Xiamen University. He is the author of more than 100 articles. He was a Visiting Scholar with the Department of Biochemistry, University of Cambridge, from 2009 to 2010, the Department of Chemistry, Hongkong Baptist University, in 2011, and the School of Medicine, University of Washington, from 2018 to 2019. His research interests include network medicine, machine learning with graph, and data analysis in biomedicine.