Graphical Modeling of Multiple Biological Pathways in Genomic Studies

Yujing Cao, Yu Zhang, Xinlei Wang, and Min Chen

1 Introduction

Many complex diseases, like various types of cancer, type 2 diabetes, and psychiatric disorders, are known to be associated with a number of genetic factors and gene expression profiles. However, the current treatments of complex diseases often fail to work well for all patients. Some patients may respond differently to the same treatment, and may suffer from adverse side effects differently. The genome and transcriptome of an individual may affect the susceptibility to develop a disease and the variation in the responses to treatments. Identifying genomic and transcriptomic risk factors thus can help us to better understand the pathogenesis of a disease. It is also the very first step toward the development of successful prevention and intervention strategies. In addition, it may shed light on genomic and transcriptomic markers that may aid the decisions of precision medicine to improve the treatment efficiency and reduce the side effects. Here, we focus on developing a

Yujing Cao and Yu Zhang authors contributed equally.

Y. Cao · Y. Zhang
Department of Mathematical Sciences, University of Texas at Dallas, Dallas, TX, USA

X. Wang
Department of Statistical Science, Southern Methodist University, Richardson, TX, USA

M. Chen (✉)
Department of Mathematical Sciences, University of Texas at Dallas, Dallas, TX, USA

Department of Population and Data Sciences, UT Southwestern Medical Center, Richardson, TX, USA
e-mail: mchen@utdallas.edu

Y. Zhao, (Din) D.-G. Chen (eds.), *Modern Statistical Methods for Health Research*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-030-72437-5_19

novel statistical model to identify the genetic factors and genes that are associated with complex diseases.

Genome-wide association study (GWAS) is a popular approach to identifying genetic variants related to a complex disease [8, 18, 40]. Single nucleotide polymorphism (SNP) is among the most common types of genetic variations and is the basic unit of investigation in many GWASs. There are millions of SNPs in the human genome, accounting for a large portion of the genomic elements that affect many phenotypes. According to the NHGRI GWAS Catalog [7], GWASs have identified over 171,000 associations as candidate risk factors in complex diseases by the year 2020. Although the traditional GWAS methods based on single-marker analysis have been successful, they fail to account for much of the heritability of phenotypes. One limitation is related to multiple testing correction, which is required to control the overall type I error when testing a large number of hypotheses. To improve the power of detection, researchers have developed pathway-based approaches in GWAS, which allow one to take into account multiple genetic variants in different loci that interact with one another. The pathway-based analysis offers an opportunity to collectively evaluate genetic variants so that the dependence among themselves can be considered in the model. Also, pathway-based studies can include markers whose effects are small and thus are hard to detect through traditional single-marker tests.

A number of pathway-based approaches have been proposed in the literature. For example, Luo et al. [30] proposed a two-stage (gene and pathway) GWAS. Chen et al. [12] proposed a supervised principal component analysis [2] to test the association between a group of SNPs and variation in disease outcome. For association tests with pathways in the presence of both common and rare variants, Pan et al. [37] extended the sum of powered score tests [36], originally developed for analysis of rare variants, to a pathway-based test that is data-adaptive at both the gene and the SNP levels. Note that the widely used kernel machine tests [48, 49] can be regarded as special cases of the sum of powered score tests. ICSNPathway [45] was developed to identify candidate causal SNPs and their corresponding candidate causal pathways. This approach integrates linkage disequilibrium analysis, functional SNP annotation, and pathway-based analysis.

Other than GWAS, gene expression analysis has been widely employed to identify genes associated with disease. Gene expression measures the expression level of mRNA that is related to the protein abundance. The regulation variation of gene expression plays a key role in shaping phenotypic differences among individuals, and as a result, it is very likely to influence disease susceptibility [13, 34]. For example, gene expression profiles from cancer and normal cells are used for comparison and reveal new disease entities [39]. Also, the involvements of gene expression are found in risk loci of the inflammatory bowel disease [32]. Single-gene based analysis has the limitation similar to that of the GWAS, namely, the lack of statistical power when a large number of hypotheses are being tested simultaneously. Many studies have proposed to incorporate the topological structures of biological pathways with gene expression data to identify differentially expressed genes. Similar to disease association status of genes, the genes that interact with

others tend to have similar expression status (differentially expressed and equally expressed) as well. Zhi et al. [51] used a discrete MRF to model the dependency of the differential expression patterns of genes in the network. Some researchers have extended the transitional enrichment analysis to topology-based enrichment approaches to identify pathways or gene sets that are significantly enriched with differentially expressed genes. Signaling pathway impact analysis (SPIA) [46] integrates the evidence of differentially expressed genes and topology structure of a signaling pathway. PathNet [15] is another enrichment method considering topology information of biological pathways.

There are many types of pathways, and most well-known ones include metabolic, gene regulatory, and signal transduction pathways. An example of a biological pathway is shown in Fig. 1. Metabolic pathways [50] are mainly concerned with a series of biochemical reactions, especially the chemical modification of the small molecule substrates of enzymes. For example, glucose is broken apart during cellular respiration to produce adenosine triphosphate (ATP), which is an energy source for the cell's functions. Gene-regulatory pathways control what genes are expressed and the expression levels of mRNA and proteins. Signal transduction pathways [25] transmit signals from cell's exterior to its interior. For instance, a chemical signal from outside the cell might direct the cell to produce protein inside the cell. Different pathways work together properly so that human body can function well and stay healthy. Much knowledge about biological pathways has been accumulated over the past decades. Consequently, a number of online resources for biological pathways are available. These knowledge bases are extensive, including

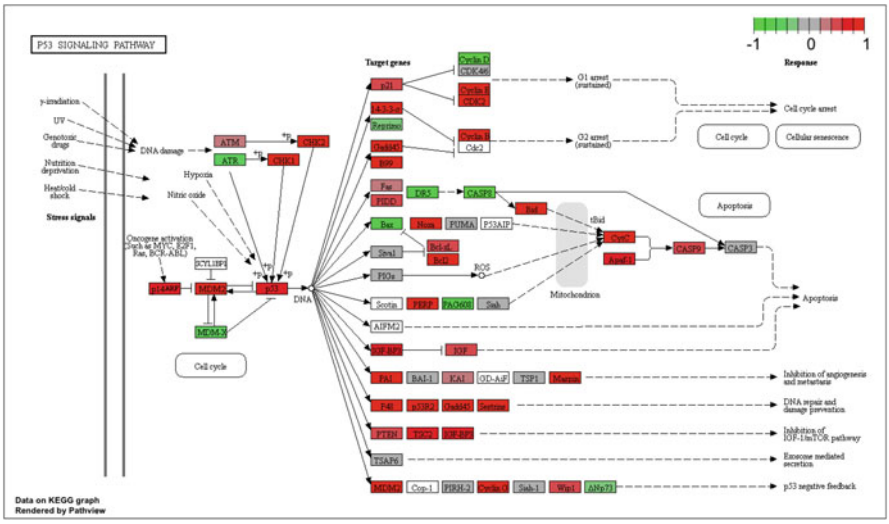


Fig. 1 p53 signaling pathway obtained from KEGG. The pathway shows the various genes, gene products, the interactions between genes, the directions of the signal propagation, and many other things

Kyoto Encyclopedia of Genes and Genomes (KEGG) [23, 24], WikiPathways [44], Reactome [21], and Pathway Commons [41].

The success of pathway-based approaches has been demonstrated in different studies. However, most pathway-based approaches utilize only partial information, that is, a pathway is treated as a list of genes and its topology structure is not considered. Indeed, a biological pathway describes a collection of interactions of molecules in cells, like mRNA, proteins, and metabolites, which coordinate with one another to perform cell functions or to direct cell responses to environmental changes. The topological structure of a biological pathway can be very informative. It reveals the interactions between genes, and it can help to improve the power of detection and to better understand risk factors of the disease. Different studies have demonstrated that incorporating the topological structure of a single biological pathway can improve the power to detect disease-related genes [10, 19, 22, 51]. Chen et al. [10] showed that the neighboring genes tend to have similar disease association statues. The proposed method introduced a Markov random field to model the topology structures of biological pathways. Hou et al. [20] proposed a novel guilt-by-rewiring principle, utilizing network information to prioritize disease genes. Freytag et al. [17] extended a logistic kernel machine into a network-based kernel machine test so that the topology structure of a biological pathway can be included in the model. Liu et al. [28] proposed the partial neighborhood selection (PNS) algorithm to estimate the gene dependence network, and a hidden Markov random field (HMRF) was adopted to combine the estimated network with genetic association scores.

The aforementioned studies only consider a single biological pathway. However, a single biological pathway only contains partial information about genes and interactions among genes. Genes participate in various biological processes simultaneously and they can interact in many different ways. A pathway usually describes very specific biological functions. As a result, genes, especially important ones, tend to interact with each other in several pathways. Therefore, combining multiple pathways can provide a more complete graph of the gene–gene interactions. For instance, genes IL23A and IL23R interact with each other in the Inflammatory Bowel Disease Pathway and Jak-STAT Signaling Pathway. Integration of these pathways can reveal and reinforce the effects of critical gene–gene interactions that play key roles in these pathways. The question arises as to whether or not we can further improve the detection power via consideration of multiple biological pathways simultaneously. Not much effort has been devoted to this important problem, although a very limited number of previous studies have shown the success of integrating multiple biological pathways. Wei and Pan [47] proposed a method to incorporate multiple gene networks, e.g., co-expression networks and functional coupling networks, with diverse genomic data to identify target genes of a transcription factor. They used a Markov random field-based mixture joint model (MRF-MJM) to merge gene networks. They assumed that the contribution of each gene network is additive and that a weight is assigned to each individual network. A larger weight indicates that there are more neighboring genes with similar status. In this method, the way to utilize multiple biological pathways is to sum over the contribution of each gene network. Their method focuses on

identifying the regulatory target genes of a transcription factor. Bokanizad et al. [6] considered another approach to combining different biological pathways. Multiple biological pathways are linked together through a single gene, called interface gene, that connects two biological pathways through biological interactions and signal transduction.

Here, we propose to combine multiple biological pathways based on the common genes shared among different biological pathways. When we merge two or more biological pathways, the topological structures of the pathways will be preserved. Also, combining different biological pathways based on the common genes they share can account for the interactions among pathways. To model the topological structures of pathways, a probabilistic graphical model called Markov random field [10, 33] is employed. An MRF is a probabilistic measure assigned to an undirected graph. In the graph, genes are nodes and interactions are denoted by edges. One advantage of MRF is that it has the ability to capture the conditional independence among variables based on the graph topology. Thus, it can provide a compact and natural representation of the joint probability distribution of the set of variables in the graph. Another advantage of the MRF is that it can be used to control the false discovery rate in the presence of dependent relationships between genes [27]. Since MRF is capable of modeling the dependent structure in data, it has been applied to a wide range of fields. For example, Lin et al. [26] estimated the differentially expressed genes in the mouse transcriptome data, using a Markov random field to model the layer similarity, temporal dependency, and the similarity between sexes.

The rest of this chapter is organized as follows. Section 2 introduces our proposed methods. It includes the basic concepts that are relevant to the graph theory, a Gibbs measure assigned to the graph serving as a prior probability, different ways of setting weights to nodes and edges, the likelihood function, and the computational method. Simulation studies are presented in Sect. 3. A small-size graph and a relatively large graph are employed to show that combining multiple biological pathways can further improve the power of detection and control the false positive rate. Section 4 shows a case study that uses lung cancer data to demonstrate the performances of the proposed methods. Finally, Sect. 5 summarizes our methods with a discussion and possible future work.

2 Method

2.1 MRF Modeling of Biological Pathways

Undirected Graphs and Biological Pathways

A biological pathway consists of a collection of interacting molecules, which can be modeled as a graph through the use of graph theory [3, 38]. We will start with introducing some basic concepts in graph theory. A graph, defined as $G = (\mathcal{V}, \mathcal{E})$, is a collection of nodes that are connected by edges, where

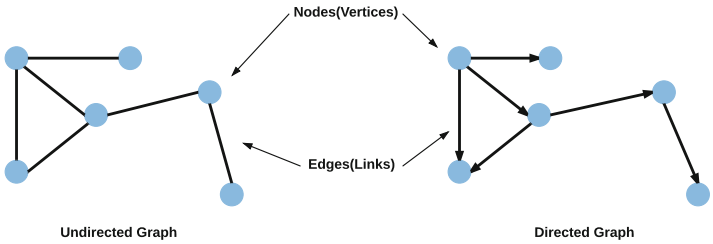


Fig. 2 Undirected graph and directed graph

$$\mathcal{V} = \{1, 2, \ldots, n\}$$
$$\mathcal{E} = \{< i, j >: i \text{ and } j \text{ are directly connected}\}.$$

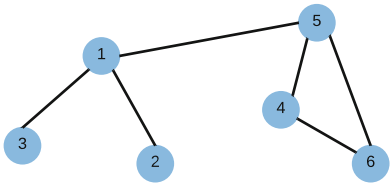
If the edges do not have directions, the graph is called an undirected graph; otherwise, it is a directed graph (see Fig. 2). There are three key concepts in graph theory that will be useful to describe the structure of a graph. The first is the neighborhood of a node v, which, by definition, is a subset of all nodes that are directly connected to v by an edge. For the ith node in $\mathcal{V}$, we define the following terms:

$$N_i = \{j :< i, j >\in \mathcal{E}\},$$
$$d_i = |N_i|,$$
$$E_{ij} = \text{the number of edges connecting node } i \text{ and node } j,$$
$$E_i = \sum_{j \in N_i} E_{ij},$$

where N_i is a set of neighbors of node i, d_i is the number of neighbors that node i has, E_{ij} is the number of edges linking node i and node j, and E_i is the number of total edges of node i. Note that $d_i = E_i$ if only one edge is present between any pair of nodes. In general, $d_i \leq E_i$ because multiple edges are allowed between node i and any of its neighbors, like in a combined graph to be discussed in Sect. 2.2. A complete graph is a simple undirected graph in which every pair of distinct nodes is connected by a unique edge. A clique is a complete subgraph of G. A clique of size k is called a k-clique (kth order clique). Each individual node is corresponding to a 1-clique. A pair of nodes can form a 2-clique and all triangles are 3-cliques. Examples are given in Fig. 3. In Fig. 3, there are six 1-cliques $\{1, 2, 3, 4, 5, 6\}$, six 2-cliques $\{(1, 2), (1, 3), (3, 6), (4, 5), (4, 6)\}$, and one 3-clique $\{(4, 6, 5)\}$. Note that since $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges, from the perspective of cliques, $\mathcal{V}$ and $\mathcal{E}$ can also be treated as a set of 1-cliques and a set of 2-cliques, respectively.

Biological pathways represent the biological reaction and interaction network in a cell. Genes, proteins, and other molecules are involved and interact with each

Fig. 3 Cliques



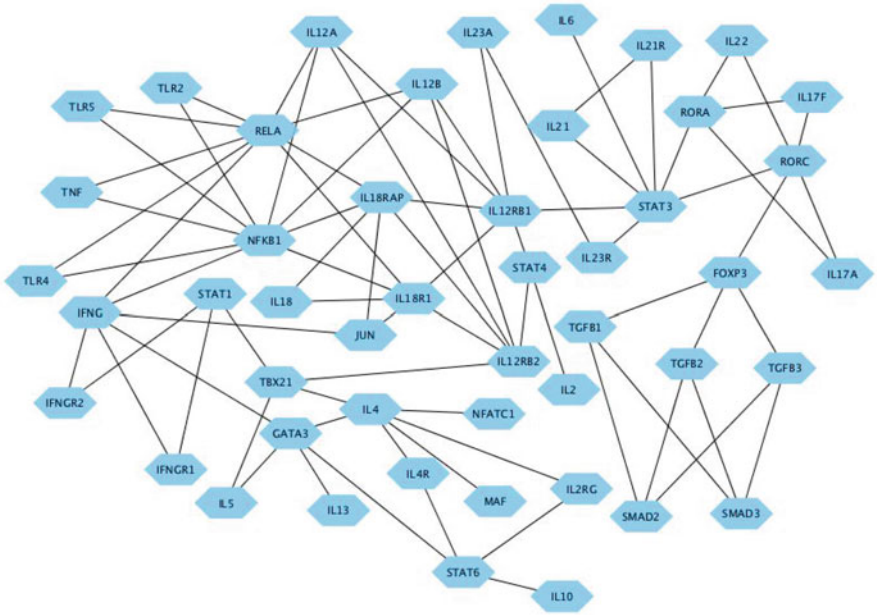


Fig. 4 Inflammatory Bowel Disease Pathway is represented as an UG (image is generated by Cytoscape [42])

other in a biological pathway. In our study, only gene–gene interactions are taken into consideration and we use an undirected graph (UG) to represent a biological pathway. An example of such a graph is shown in Fig. 4.

In the graph, genes are treated as nodes and gene–gene interactions are edges. Define S_i as the true status of gene i:

$S_i = +1$ if gene i is associated with disease or is differentially expressed,

$S_i = -1$ if gene i is not associated with disease or is not differentially expressed.

Hereafter, ± 1 are referred to as labels of nodes. Let $\mathbf{S} = (S_1, \ldots, S_n)$ be the labeling of $\mathcal{V}$. Thus, $\mathbf{S}$ is a spatial random vector whose element may be correlated to each

other. Each node can be labeled as either $+1$ or -1. So, there are 2^n unique labelings of the graph, and each unique labeling of the graph is also called a configuration. The ultimate goal is to infer the value of $\mathbf{S}$ based on the underlying topological structures of biological pathways and observed data from biological experiments.

In practice, genes do not function in isolation. For complex diseases, multiple genes have been identified to collectively account for clinical phenotypes [29]. Moreover, the same pair of genes can interact in different biological pathways, which motivates us to combine multiple biological pathways to gather more information about gene–gene interactions. Let $\mathcal{P}$ denote a set of g distinct biological pathways:

$$\mathcal{P} = \{P_1, P_2, \ldots, P_g\},$$

where $P_l = (\mathcal{V}_l, \mathcal{E}_l), l = 1, \cdots, g$.

Multiple biological pathways are combined into a big pathway, which will be integrated with genomic and transcriptomic data later. We use an intuitive approach based on the overlapping genes among the pathways to combine multiple biological pathways. As we mentioned earlier, genes may appear in different biological pathways. Based on the common genes they share, these biological pathways can be combined. Fig. 5 shows an example of combining two biological pathways.

Note that some pairs of genes are linked by multiple edges in the combined graph. The number of edges denotes how many biological pathways that the corresponding

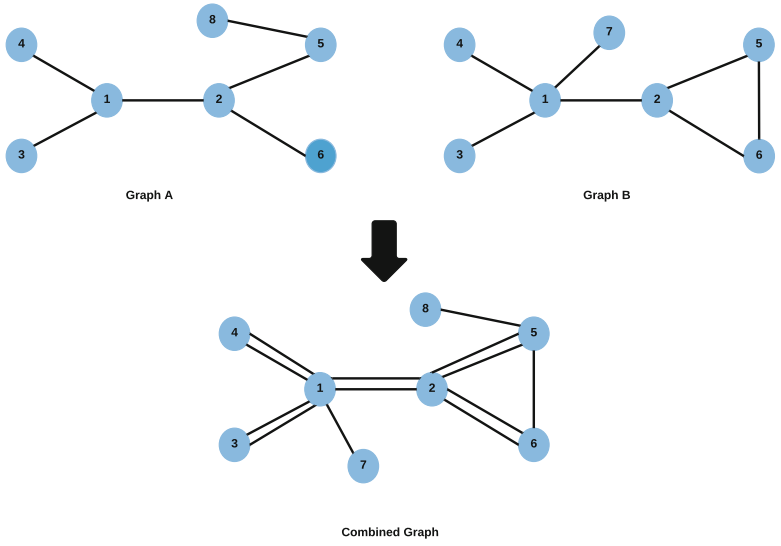


Fig. 5 An example of combining two biological pathways. Overlapping nodes in Graph A and Graph B are nodes (1, 2, 3, 4, 5, 6). Based on the shared nodes, Graphs A and B are integrated to a combined graph

genes interact with each other. In our study, we treat the multiple edges between neighboring genes as weighted single edge. The specific ways to assign weights to the weighted single edges will be discussed in section "A Nearest Gibbs Measure".

Genes interact with each other in the biological pathway. In the other words, there exist dependencies among the genes. In order to describe the topological structure and the dependencies, a Markov random field is employed [10], as will be introduced in the following section. An MRF not only describes the structures of biological pathways but also allows us to define a joint distribution for interdependent genes.

A Nearest Gibbs Measure

Speaking about the joint distribution of genes, we need to assign a probability measure to the combined biological pathway. The probability measure should reflect that the neighboring genes tend to have similar labels, and it should also quantify the effects of edges that connect the neighboring genes. Following [10], we can achieve both goals with one probability measure, a nearest Gibbs measure, as follows:

$$\mathbb{P}(\mathbf{S}|\theta_0) = \frac{1}{z(\theta_0)} \exp\left\{ h \sum_{i\in\mathcal{V}} I_1(S_i) + \tau_0 \sum_{<i,j>\in\mathcal{E}} (\omega_i + \omega_j) I_{-1}(S_i) I_{-1}(S_j) \right.$$
$$\left. + \tau_1 \sum_{<i,j>\in\mathcal{E}} (\omega_i + \omega_j) I_1(S_i) I_1(S_j) \right\}, \tag{1}$$

where $\theta_0 = (h, \tau_1, \tau_0)$, $I_1(\cdot)$ and $I_{-1}(\cdot)$ are indicator functions, and $z(\theta_0)$ is a normalizing function that is the sum over all 2^n possible configurations:

$$z(\theta_0) = \sum_{\mathbf{S}} \exp\left\{ h \sum_{i\in\mathcal{V}} I_1(S_i) + \tau_0 \sum_{<i,j>\in\mathcal{E}} (\omega_i + \omega_j) I_{-1}(S_i) I_{-1}(S_j) \right.$$
$$\left. + \tau_1 \sum_{<i,j>\in\mathcal{E}} (\omega_i + \omega_j) I_1(S_i) I_1(S_j) \right\}. \tag{2}$$

Note that it is computationally prohibitive to evaluate $z(\theta_0)$ when n is large. For instance, there are over 10 billion possible configurations when a graph has 30 nodes. Here, τ_0 and τ_1 assign weights to the 2-cliques in which both nodes are negative and positive genes, respectively; ω_i is a function of d_i and E_i, reflecting a weight we assign to node i. The details of the function ω_i will be described later in Sect. 2.2 when we define methods of combining biological pathways.

The probability measure in (1) directly considers the topological structure of a pathway. The first term is the sum over all the 1-cliques; the second sum and the

third sum are taken over all the 2-cliques that contain both of two nodes labeled as -1 and labeled as $+1$, respectively. Positive τ_0 and τ_1 will put more weights on the 2-cliques in which all of the included nodes have the same labels, which is desirable in our context. The parameter h determines the marginal probability of S_i when $\tau_0 = \tau_1 = 0$, i.e., if all nodes are isolated, which indicates that they are independent:

$$\mathbb{P}(S_i = 1|h, \tau_0 = \tau_1 = 0) = \frac{\exp(h)}{\exp(h) + 1}.$$

There is an attractive feature of the Gibbs measure, that is, a sample from Gibbs measure has the local Markov property. This property defines an MRF on S, which by definition is $Pr(S_i|S_{v-i}) = Pr(S_i|S_{N_i})$, where $\mathcal{V} - i$ denotes all the nodes but i, and N_i is the set of all immediate neighbors of node i. This property can be asserted by the Hammersley–Clifford theorem [5]. We use an MRF to model the interactions between genes that are directly linked.

Theorem 1 (Hammersley–Clifford Theorem) *The spatial random vector,* **S***, under the Gibbs measure, is a Markov random field and thus satisfies*

$$\mathbb{P}(S_i|S_{v-i}, \theta_0) = \mathbb{P}(S_i|S_{N_i}, \theta_0).$$

Also, the conditional distribution of an MRF has a logistic regression form as shown below [10]:

$$\begin{aligned}\text{logit}(\mathbb{P}(S_i|S_{N_i}, \theta_0)) = h - \tau_0 \sum_{<i,j>\in\mathcal{E}} (\omega_i + \omega_j) I_{-1}(S_i) I_{-1}(S_j) \\ +\tau_1 \sum_{<i,j>\in\mathcal{E}} (\omega_i + \omega_j) I_1(S_i) I_1(S_j), \quad i = 1, \ldots, n. \quad (3)\end{aligned}$$

Equivalently, (3) can be written as a system of linear equations:

$$\begin{aligned}\text{logit}(\mathbb{P}(S_i|S_{N_i}, \theta_0)) = \beta_{i0} + \beta_{i1} S_1 + \cdots + \beta_{in} S_n, \\ i = 1, \ldots, n, \quad (4)\end{aligned}$$

where

$$\beta_{i0} = h$$

$$\beta_{ij} = \begin{cases} 0 & \text{if } i = j \text{ or } <i, j> \notin \mathcal{E} \\ (\omega_i + \omega_j)\{\tau_1 I_1(S_j) + \tau_0 I_{-1}(S_j)\} & \text{if } <i, j> \in \mathcal{E}. \end{cases}$$

The Markov property implies that the conditional distribution of S_i, given all the other node labels in the network, is equivalent to the conditional distribution of

S_i given all its immediate neighbors. If S_i and S_j are not neighbors, then they are conditionally independent. Now, we give an interpretation of ω_i in (1). From (4), it is obvious that the conditional probability of S_i depends on the weighted sum of its neighbors. Moreover, S_i has different weights depending on the sizes of cliques used to describe the structure of the graph in the probability measure.

2.2 Combine Multiple Pathways

In Eq. (1), the weight of S_i is

$$\begin{aligned}(\omega_i + \omega_j)\tau_1 \quad & \text{if } S_i = S_j = +1,\\ (\omega_i + \omega_j)\tau_0 \quad & \text{if } S_i = S_j = -1.\end{aligned}$$

Here, $(\omega_i + \omega_j)$ is the sum of weights over all the nodes in the same 2-cliques. Recall that in the combined graph shown in Fig. 5, a pair of nodes can be linked by more than one edge, which may indicate the strength of relation between the neighboring nodes. The weights of nodes and edges are related to the number of neighbors and edges that nodes have in the combined graph. Next, we will present four different probability measures in which ω_i and the weights of edges are set in different ways.

Method 1 In this method, if two or more edges are between two neighboring genes, we only count them once. We set ω_i to be the logarithm of d_i, the number of neighbors of S_i. As a result, a gene that interacts with many other genes in the pathway has a large weight because it may play a central role in a biological process, and thus it is likely to have a large influence. However, a gene with one neighbor is assigned with 0. Thus, its effect has been reduced. The probability measure and the logistic form of the first method are identical to the equations shown in (1) and (3), respectively.

In a combined graph, if multiple edges are present between a pair of nodes, one could assign a weight to this link, in addition to the weights $(\omega_i + \omega_j)$ that are based on the nodes. In Methods 2 through 4 below, we define the weight of the link between nodes i and j as $(E_{ij})^2 \cdot (AE/TE)$, where E_{ij} is the number of edges linking nodes i and j,

$$AE = \frac{\sum_{<i,'j'>\in\mathcal{E}_1} E^2_{i'j'} + \cdots + \sum_{<i,'j'>\in\mathcal{E}_g} E^2_{i'j'}}{g},$$

$$TE = \sum_{<i,'j'>\in\mathcal{E}_{CP}} E^2_{i'j'},\ \mathcal{E}_{CP} \text{ denotes the edge set of the combined pathway.}$$

Note that E_{ij}, the number of edges between two nodes in the combined pathway, never decreases as more pathways are added. To regularize the growth of the edge

weights, we multiply $(E_{ij})^2$ by a normalizing factor (AE/TE). The probability measure of $\mathbf{S}$ thus becomes

$$\mathbb{P}(\mathbf{S}|\theta_0) = \frac{1}{z(\theta_0)} \exp\left\{ h \sum_{i \in \mathcal{V}_{CP}} I_1(S_i) + \tau_0 \sum_{<i,j> \in \mathcal{E}_{CP}} (\omega_i + \omega_j) I_{-1}(S_i) I_{-1}(S_j) E_{ij}^2 \left(\frac{AE}{TE}\right) \right.$$
$$\left. + \tau_1 \sum_{<i,j> \in \mathcal{E}_{CP}} (\omega_i + \omega_j) I_1(S_i) I_1(S_j) E_{ij}^2 \left(\frac{AE}{TE}\right) \right\}. \tag{5}$$

The above probability measure also defines an MRF. The corresponding logistic form is

$$\begin{aligned} \text{logit}(\mathbb{P}(S_i | S_{N_i}, \theta_0)) = h - \tau_0 & \sum_{<i,j> \in \mathcal{E}_{CP}} (\omega_i + \omega_j) I_{-1}(S_i) I_{-1}(S_j) E_{ij}^2 \left(\frac{AE}{TE}\right) \\ & + \tau_1 \sum_{<i,j> \in \mathcal{E}_{CP}} (\omega_i + \omega_j) I_1(S_i) I_1(S_j) E_{ij}^2 \left(\frac{AE}{TE}\right), \\ & i = 1, \ldots, n. \end{aligned} \tag{6}$$

The system of linear equations of (6) is

$$\begin{aligned} \text{logit}(\mathbb{P}(S_i | S_{N_i}, \theta_0)) = \beta_{i0} + \beta_{i1} S_1 + \cdots + \beta_{in} S_n, \\ i = 1, \ldots, n, \end{aligned} \tag{7}$$

where

$$\beta_{i0} = h$$

$$\beta_{ij} = \begin{cases} 0 & \text{if } i = j \text{ or } <i, j> \notin \mathcal{E}_{CP} \\ (\omega_i + \omega_j)\{\tau_1 I_1(S_j) + \tau_0 I_{-1}(S_j)\} E_{ij}^2 (\frac{AE}{TE}) & \text{if } <i, j> \in \mathcal{E}_{CP}. \end{cases}$$

Methods 2 through 4 differ in the definition of ω_i, the weight assigned to node i. The motivation is to give more credit to the nodes that have more neighbors or more total number of edges in the combined graph.

Method 2 $\omega_i = \log(\frac{E_i}{g})$, where E_i is the total number of edges of node i.

Method 3 $\omega_i = \log(d_i)$, where d_i is the size of neighborhood of node i.

Method 4 $\omega_i = \log(E_i)$.

The probability measures in (1) and (5) both define an MRF that can be applied to describe the pathway topology. The MRF will be treated as a prior distribution

under a Bayesian model to help us integrate the topological structure of biological pathways and prior biology knowledge in the Bayesian framework later (the details of Bayesian framework will be shown in Sect. 2.4).

2.3 Likelihood Function

We follow the method proposed by Chen et al. [10] to form a likelihood function. The evidence about disease association status or DE status, which is gathered from biological experiments, can be summarized by p-values at gene level. For gene i, its p-value can be converted to a response variable y_i through

$$y_i = \Phi^{-1}(1 - p_i),$$

where p_i is the p-value and $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0, 1)$. Note that a small value of p_i corresponds to a large value of y_i. Assume y_i are conditionally independent given S, the status of all genes. The null hypothesis is that the gene is unrelated to the disease. Under the null case where $S_i = -1$, the distribution of y_i is standard normal distribution. Therefore, the density of y_i is $f_0(y_i) \sim \mathcal{N}(0, 1)$. When the alternative hypothesis is true, that is, $S = +1$, the distribution of y_i is assumed to follow a normal distribution with the mean μ_i and the variance σ_i^2, where μ_i and σ_i are unknown. To account for the variations of μ_i and σ_i, prior distributions need to be assigned to μ_i and σ_i. We employ conjugate priors $\mu_i|\sigma_i^2 \sim \mathcal{N}(\bar{\mu}, \frac{\sigma_i^2}{a})$ and $\sigma_i^2 \sim$ Inverse Gamma$(\frac{\nu}{2}, \frac{\nu d}{2})$ for efficient computations. Define $\theta_1 = (\bar{\mu}, a, \nu, d)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$. Under this prior setting, the marginal density of y_i is

$$\begin{aligned}
f_1(y_i|S_i = 1, \theta_1) = \int\int \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp[\frac{-(y_i - \mu_i)^2}{2\sigma_i^2}] \frac{\sqrt{a}}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{-a(y_i - \bar{\mu})^2}{2\sigma_i^2}\right] \\
\times \frac{(\nu d/2)^{\nu/2}}{\Gamma(\nu/2)} (\sigma_i^{-2})^{\nu/2-1} \exp\left[-(\sigma_i^{-2})\frac{\nu d}{2}\right] d\mu_i d(\sigma_i^{-2}) \\
= \int \frac{1}{\sqrt{2\pi\frac{a+1}{a}\sigma_i^2}} \exp[\frac{-a(y_i - \bar{\mu})^2}{2(a+1)\sigma_i^2}] \frac{(\nu d/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} (\sigma_i^{-2})^{\nu/2-1} \exp\left[\frac{-\nu d\sigma_i^{-2}}{2}\right] d(\sigma_i^{-2}) \\
= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{a}}{\sqrt{a+1}} \frac{(\nu d/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} \int (\sigma_i^{-2})^{\frac{\nu+1}{2}} \exp\left[-\sigma_i^2\{\frac{\nu d}{2} + \frac{a(y_i - \bar{\mu})^2}{2(a+1)}\}\right] d(\sigma_i^{-2}) \\
= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{a}}{\sqrt{a+1}} \frac{(\nu d/2)^{\nu/2}}{\Gamma(\nu/2)} \Gamma(\frac{\nu+1}{2}) \cdot \{\frac{\nu d}{2} + \frac{1}{2}\frac{a}{a+1}(y - \bar{\mu})^2\}^{-\frac{(1+\nu)}{2}} \\
= \pi^{-1/2}(\nu d)^{\nu/2} \frac{\sqrt{a}}{\sqrt{a+1}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(\frac{a}{a+1}(y_i - \bar{\mu}^2 + \nu d)\right)^{-(1+\nu)/2}. \qquad (8)
\end{aligned}$$

Therefore, the likelihood function of $\boldsymbol{y}$ is

$$f(\boldsymbol{y}|\mathbf{S}, \theta_1) = \prod_{j:S_j=-1} f_0(y_j) \times \prod_{j:S_j=+1} f_1(y_j|S_j = 1, \theta_1). \tag{9}$$

2.4 Posterior Probability Under Bayesian Framework

Under a Bayesian framework, the MRF is the prior probability describing the topological structures of the biological pathways, and the likelihood function in Eq. (9) presents the evidence from biological experiments. Thus, the posterior probability of $\mathbf{S}$, given the observed data $\boldsymbol{y}$, is

$$\begin{aligned} \mathbb{P}(\mathbf{S}|\boldsymbol{y}, \theta_0, \theta_1) &= \frac{f(\boldsymbol{y}|\mathbf{S}, \theta_1)\mathbb{P}(\mathbf{S}|\theta_0)}{\sum_{\mathbf{S}} f(\boldsymbol{y}|\mathbf{S}, \theta_1)\mathbb{P}(\mathbf{S}|\theta_0)} \\ &\propto f(\boldsymbol{y}|\mathbf{S}, \theta_1)\mathbb{P}(\mathbf{S}|\theta_0). \end{aligned} \tag{10}$$

Recall that the prior probability introduced in Sect. 2.1 defines an MRF and has a nice conditional probability. Similar to the prior probability, the posterior probability defines an MRF as well. For node i,

$$\begin{aligned} \mathbb{P}(S_i = +1|\boldsymbol{y}, S_{v-i}, \theta_0, \theta_1) &\propto f_1(y_i|\theta_1)\mathbb{P}(S_i = +1|S_{v-i}, \theta_0) \\ &= f_1(y_i|\theta_1)\mathbb{P}(S_i = +1|S_{N_i}, \theta_0). \end{aligned} \tag{11}$$

The conditional posterior distribution of S_i, given all other nodes, only depends on its neighbors, which means that the posterior distribution leads to an MRF [10]. We use method 1 as an example to show the logistic form of the conditional distribution of S_i:

$$\begin{aligned} \text{logit}(\mathbb{P}(S_i|\boldsymbol{y}, S_{N_i}, \theta_0, \theta_1) = h + \log LR(y_i; \theta_1) - \tau_0 \sum_{<i,j>\in\mathcal{E}_{CP}} (\omega_i + \omega_j) I_{-1}(S_i) I_{-1}(S_j) \\ + \tau_1 \sum_{<i,j>\in\mathcal{E}_{CP}} (\omega_i + \omega_j) I_1(S_i) I_1(S_j), \end{aligned} \tag{12}$$

where

$$LR(y_i; \theta_1) = \frac{f_1(y_i|\theta_1)}{f_0(y_i)},$$

the marginal likelihood ratio. Therefore, (12) integrates the evidence from biological experiments that is reflected by the marginal likelihood ratio and the effect from

interactions among neighboring genes in biological pathway reflected by the conditional prior odds. It is easy to see that the posterior conditional logit form in (12) is the same as the prior conditional logit in (3) except its intercept is $h + \log(LR(y_i); \theta_1)$. The observed log-likelihood ratio provides an additive effect to the logit of prior.

We can also rewrite (12) in the form of a system of linear regressions:

$$\begin{aligned} \text{logit}(\mathbb{P}(S_i|\boldsymbol{y}, S_{N_i}, \theta_0, \theta_1)) &= \beta_{i0} + \beta_{i1}S_1 + \cdots + \beta_{in}S_n, \\ & i = 1, \ldots, n, \end{aligned} \tag{13}$$

where

$$\beta_{i0} = h + \log LR(y_i; \theta_1),$$

$$\beta_{ij} = \begin{cases} 0 & \text{if } i = j \text{ or } < i, j > \notin \mathcal{E}_{CP}, \\ (\omega_i + \omega_j)\{\tau_1 I_1(S_j) + \tau_0 I_{-1}(S_j)\} & \text{if } < i, j > \in \mathcal{E}_{CP}. \end{cases}$$

The posterior probabilities of other three priors proposed in Sect. 2.1 have similar logistic regression forms. As mentioned before, the differences among Method 2, Method 3, and Method 4 are the definitions of ω_i.

2.5 Monte Carlo Markov Chain (MCMC) Simulation

As the number of genes becomes very large, it is prohibitive to calculate the posterior probability directly. But the posterior distribution has a nice closed-form conditional distribution, due to the Markov property. It is easier to sample from the conditional distribution using the Gibbs sampling [9]. The Gibbs sampler is one of the MCMC algorithms, and it can generate a sequence of samples from the conditional distributions.

In our context, the specific steps are described as the following: we start by setting the initial values of $\boldsymbol{S}$, $\boldsymbol{s}^{(0)} = (s_1, \ldots, s_n)$. Here, the upper case $\boldsymbol{S}$ is to denote a random vector and use the lower case $\boldsymbol{s}^{(k)}$ to denote a realization of the random vector in the kth iteration. The elements of the vector $\boldsymbol{s}^{(k)}$ are $+1$s and -1s. At iteration k, we update the labels sequentially for $i = 1, \ldots, n$ based on

$$\begin{aligned} &\text{logit}(\mathbb{P}(s_i^{(k)}|\boldsymbol{y}, s_1^{(k)}, \cdots, s_{i-1}^{(k)}, s_i^{(k-1)}, \cdots, s_n^{(k-1)}, \theta_1, \theta_0) \\ &\quad = \beta_{i0} + \beta_{i1}s_1^{(k)} + \cdots + \beta_{i,i-1}s_{i-1}^{(k)} + \beta_{i,i+1}s_{i+1}^{(k-1)} + \cdots + \beta_{in}s_n^{(k-1)}. \end{aligned}$$

When performing the Gibbs sampling, we recommend to restart the simulation multiple times with different initial values to reduce the influence of initial values and ignore a number of simples from beginning (the so-called burn-in period).

2.6 Making Inference Based on the Marginal Posterior Probability

In GWAS and mRNA expression studies, a set of genes is identified as candidates that are very likely to be associated with diseases or differentially expressed. Therefore, we want to include as many truly positive genes among the candidates as possible. Following [10], we describe a method that can be used to rank order genes. The inference of gene status is based on $m_i = \mathbb{P}(S_i = 1|\boldsymbol{y}, \theta_0, \theta_1)$, the mean of the marginal posterior probability of S_i. A decision rule in the form $\delta(m_i) = I(m_i \geq m^*)$ is considered, where $I(\cdot)$ is an indicator function and m^* is a decision threshold. Here, m^* can facilitate deciding the status of a gene as below:

$$\delta(m_i) = \begin{cases} 1 & m_i \geq m^* \\ 0 & m_i < m^*. \end{cases}$$

If $\delta(m_i)$ is 1, the decision is positive, indicating that gene i is considered to be associated with the disease or differentially expressed; otherwise, gene i is identified as a negative gene. To find the decision threshold m^*, a 0–1 loss function, which is widely used in classification problem, is employed. In our context, a 0–1 loss function is defined by $L(\boldsymbol{S}, \delta) = \sum_{i=1}^{n} |I_1(S_i) - \delta(m_i)|$. This loss function penalizes equally the false positive and false negative errors. Note that $L(\boldsymbol{S}, \delta)$ is a random variable because it is a function of the random vector $\boldsymbol{S}$ and a decision function $\delta(\cdot)$, which depends on $E[\boldsymbol{S}|y]$ and m^*. We consider the expected loss with respect to the posterior distribution of $\boldsymbol{S}$:

$$E\{L(\boldsymbol{S}, \delta)|\boldsymbol{y}, \theta_0, \theta_1\} = \sum_{i=1}^{n} |I_1(S_i) - \delta(m_i)| \cdot \mathbb{P}(S_i|\boldsymbol{y}, \theta_0, \theta_1). \tag{14}$$

Then, m^* is sought to minimize the expected loss:

$$\begin{aligned} m^* &= \underset{m}{\operatorname{argmin}}\, E\{L(\boldsymbol{S}, \delta)|\boldsymbol{y}, \theta_0, \theta_1\} \\ &= \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} |I_1(S_i) - \delta(m_i)| \cdot \mathbb{P}(S_i|\boldsymbol{y}, \theta_0, \theta_1). \end{aligned} \tag{15}$$

To find the solution, look at the loss incurred by gene i: $[1 - \delta(m_i)] \cdot m_i + \delta(m_i) \cdot (1 - m_i)$. To minimize it, $\delta(m_i)$ should be 1 if $m_i \geq 0.5$ and 0 otherwise. Therefore, the expected loss in Eq. (14) can be minimized when $m^* = 0.5$. For other possible decision rules, please see [10].

3 Simulation Studies

Two simulation studies are carried out to examine how the values of (h, τ_1, τ_0) affect the network and evaluate the performances of the proposed methods. A small combined network that only contains 10 nodes is used to explore the effect of prior settings. A relatively large combined network having 27 nodes is used to evaluate the performance of combined network based on the control of false positive rates and false discovery rates.

4 10-Node Network

A 10-node network is used to study the effects of hyper-parameters h, τ_1, and τ_0. In the network shown in Fig. 6, Graph A has 6 nodes, and Graph B has 10 nodes. Combined together, there is a total of 10 nodes in the network. Nodes (1, 2, 4, 10) are labeled as $+1$ (in red color) and nodes (5, 6, 7, 8, 9) are labeled as -1 (in blue color). Graph A and Graph B share 4 positive nodes (1, 2, 3, 4). In addition, Graph B has one more positive node (Node 10) than Graph A. After combining the two graphs based on the common nodes, we can obtain a combined graph in which multiple edges exist. Compared to the single graph A or B, some neighbors are connected by two edges. Consequently, neighboring nodes with identical labels have reinforced relationship if connected by multiple edges. The more edges the neighboring genes share, the stronger the relationship they have. As a result, it is more likely for the neighboring genes to have the same status. To conduct a fair comparison between using only Graph A or B and using combined 10-node

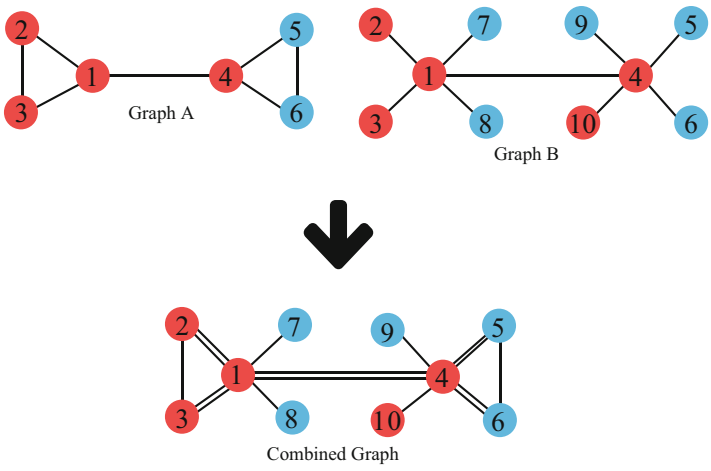


Fig. 6 Simulated 10-node networks and combined network

network, the same set of nodes has to appear in both Graphs A and B, as do in the combined graph. Therefore, the nodes in Graph B but not in Graph A are added as singletons to Graph A.

When $S = +1$, to simulate different levels of the power of the statistical tests, p-values are calculated from two-sided z-scores drawn from $\mathcal{N}(0.5, 1)$, $\mathcal{N}(1, 1)$, and $\mathcal{N}(1.5, 1)$, corresponding to the power of 0.08 (weak), 0.16 (median), and 0.32 (strong), respectively, for the tests. When $S = -1$, p-values are sampled from Uniform(0, 1).

To study the effects of hyper-parameters (h, τ_1, τ_0), Table 1 lists four main groups of prior settings. They are chosen to control the prior mean $\mathbb{P}(S_i = +1)$ to be around 0.05, 0.15, 0.25, and 0.4, respectively, for the four groups. The average values of $\mathbb{P}(S_i = +1)$ are listed in the column $\mathbb{E}[\Pr(S_i = 1)]$. Under each main group, there are two subgroups. The difference of the two subgroups is in the average probability of $\mathbb{P}(S_i = S_j = +1)$, shown in the column $\mathbb{E}[\Pr(S_i = S_j = 1)]$. The values of $(\bar{\mu}, a, \nu, d)$ in likelihood function are set to be (3,1,10,1).

One thousand datasets are simulated for every prior setting. For the 10-node network, the posterior probability can be calculated directly from the global measure without using the Gibbs sampling. This is because there are only 1024 configurations in total, and it is not computation-intensive to evaluate the normalizing term in the global measure in Eq. (10). To rank the genes based on p-values or marginal posterior $P(S_i = +1|y)$ from the proposed methods, we can calculate true positive rate and false positive rate:

$$\text{true positive rate} = \frac{\text{number of true positives}}{\text{number of true positives} \ + \ \text{number of false positives}},$$

$$\text{false positive rate} = \frac{\text{number of false positives}}{\text{number of true positives} + \ \text{number of false positives}}.$$

By varying the cutoff values, we can plot true positive rate versus false positive rate to draw the receiver operating characteristic (ROC) curve. Finally, the area under the ROC curve (AUC) can be calculated. The value of AUC is 0.5 without

Table 1 Prior settings for the 10 nodes combined network

Group	Subgroup	Parameters			Prior mean	
		h	τ_1	τ_0	$\mathbb{E}[\Pr(S_i = 1)]$	$\mathbb{E}[\Pr(S_i = S_j = 1)]$
1	a	-3.000	0.100	0.001	0.0483	0.0029
	b	-2.750	0.150	0.005	0.0616	0.0051
2	a	-2.000	0.200	0.001	0.1351	0.0275
	b	-2.000	0.250	0.005	0.1397	0.0322
3	a	-1.250	0.100	0.001	0.2430	0.0717
	b	-1.500	0.250	0.005	0.2329	0.0861
4	a	-0.500	0.050	0.005	0.3956	0.1703
	b	-1.000	0.250	0.010	0.3660	0.1965

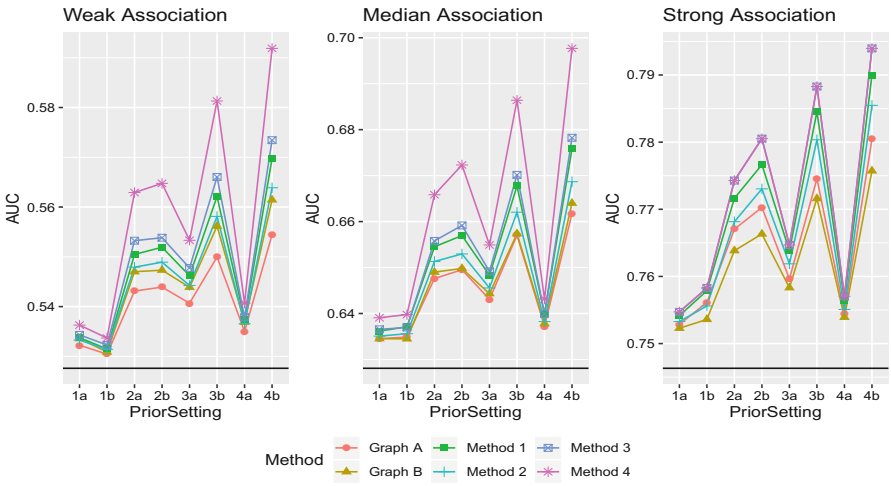


Fig. 7 AUC of the 10-node pathway. Graph A and Graph B are referring to methods based on single pathways A and B, respectively. Methods 1–4 are the proposed ones to combine both pathways. The black horizontal lines indicate AUC based on p-values only

any models, and a value of the AUC higher than 0.5 means that the performance of the model is better. The AUCs calculated using only the p-values are 0.5276, 0.6281, and 0.7463, corresponding to weak, median, and strong associations. Fig. 7 shows the performance of proposed methods.

First of all, in Fig. 7, the values of AUC from Bayesian models (Graph A, Graph B, and Methods 1–4) are larger than the values obtained using p-values alone (the black horizontal lines), no matter using a single pathway (Graph A and Graph B) or the combined one (Methods 1–4). Using the combined graph outperforms the single graph, especially in prior setting 2, prior setting 3, and prior setting 4. In general, the values of AUC that are corresponding to subgroup (b) are higher than those to subgroup (a). The reason is that (h, τ_1, τ_0) in subgroup (b) are larger than the ones in subgroup (a). So, priors in subgroup (b) encourage nodes labeled as $+1$. In general, the value of τ_1, which is the weight of linked truly associated or equally expressed genes, should be larger than τ_0.

5 27-Node Network

The proposed methods are also applied to two large networks and the combined network in Fig. 8. There are 21 nodes in Graph A and 24 nodes in Graph B. Nodes 1, 2, 3, 4, 5, 6, 19, 22, and 23 are considered as true positive genes (in red color), and the others are negative genes (in blue color). One positive node (#19) and two negative nodes (#20 and #21) in Graph A are not presented in Graph B. On the other

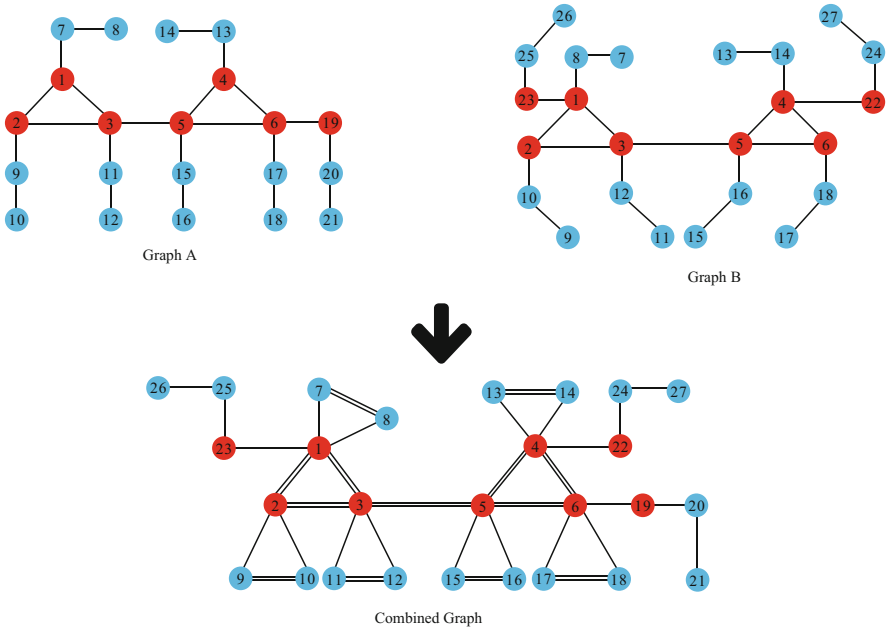


Fig. 8 Simulated 27-node networks and the combined network

Table 2 Prior settings for the combined 27-node pathway

Group	Subgroup	Parameters			Prior mean	
		h	τ_1	τ_0	$\mathbb{E}[\Pr(S_i = 1)]$	$\mathbb{E}[\Pr(S_i = S_j = 1)]$
1	a	-3.000	0.100	0.001	0.0470	0.0031
	b	-2.750	0.150	0.005	0.0626	0.0061
2	a	-2.000	0.200	0.001	0.1414	0.0308
	b	-2.000	0.250	0.005	0.1533	0.0434
3	a	-1.250	0.100	0.001	0.2491	0.0767
	b	-1.500	0.250	0.005	0.2602	0.1139
4	a	-0.500	0.050	0.005	0.3991	0.1740
	b	-1.000	0.250	0.010	0.4184	0.2686

hand, two other positive nodes (#22 and #23) and four negative ones (#24, #25, #26, and #27) in Graph B are not in Graph A. When they are combined together, we obtain a 27-node network. The prior settings are the same as that chosen for 10-node networks. Table 2 shows the average prior probabilities $\mathbb{P}(S_i = 1)$ and the average prior probabilities $\mathbb{P}(S_i = S_j = 1)$.

The same method is applied to simulate p-values, that is, they are computed from two-sided z-scores drawn at random from $\mathcal{N}(1, 1)$, $\mathcal{N}(1.5, 1)$, and $\mathcal{N}(2, 1)$

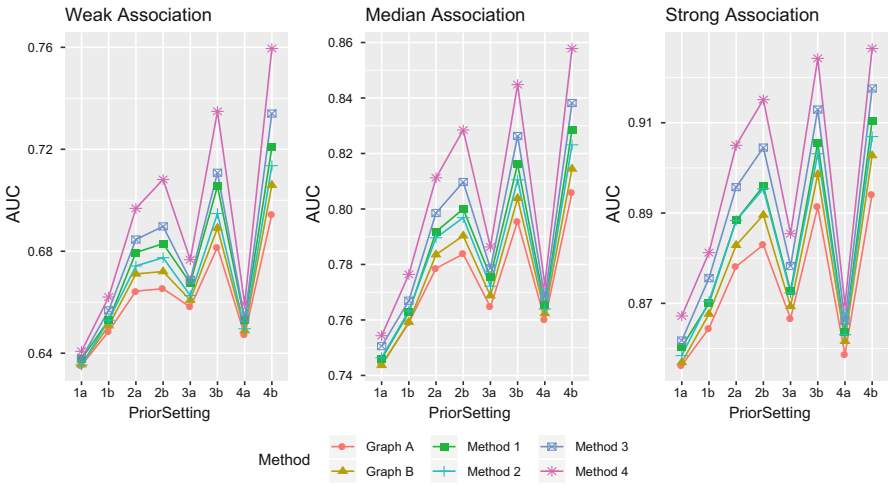


Fig. 9 AUC of the 27-node pathway. Graph A and Graph B refer to methods based on single pathways A and B, respectively. Methods 1–4 are the proposed ones to combine both pathways. The black horizontal lines indicate AUC based on p-values only

when $S = +1$, and p-values are sampled from Uniform(0,1) when $S = -1$. The AUC is computed for each group of prior settings to evaluate the performances of the proposed methods. We simulate 100 datasets and run a Gibbs sampler with 10 restarts where each restart contains 1000 iterations (the first 100 are burn-ins). For each simulated dataset, we calculate the value of AUC. Fig. 9 shows the average values of AUC of the 100 simulations for all scenarios. The values of AUC computed from p-values alone are 0.6367, 0.7489, and 0.8483, corresponding to weak, median, and strong tests, respectively. Form the figure, similar observations can be drawn as from the 10-node network.

In addition to comparisons based on the AUC, next we evaluate the performances of the proposed methods in terms of false positive rate (FPR), true positive rate (TPR), and false discovery rate (FDR). We apply the decision rule $\delta(m_i)$ to the marginal posterior probability with a cutoff $m^* = 0.5$. Table 3 lists the average FPR, TPR, and FDR of 100 datasets with 8 different prior settings. We also compare the proposed methods with the p-value method with a cutoff value of 0.05.

In Table 3, for each prior setting, the proposed methods that make use of multiple networks have higher TPR and lower or equal FDR than using a single network. For prior setting groups 1, 2, and 3, the FPR of the proposed methods is much lower than 0.05, making the TPR worse than the method of p-value only. However, the prior setting 4b controls FPR at $\sim$0.05, and it has a higher TPR and a lower FDR than using p-value alone.

Table 3 Average false positive rate (FPR), true positive rate (TPR), and false discovery rate (FDR)

Group	Method	Weak association			Median association			Strong association		
		TPR	FPR	FDR	TPR	FPR	FDR	TPR	FPR	FDR
	p-Value	0.1578	0.0528	0.4238	0.3111	0.0528	0.2567	0.5189	0.0528	0.1549
1a	Graph A	0.0489	0.0056	0.6183	0.1100	0.0056	0.3278	0.2389	0.0056	0.0979
	Graph B	0.0489	0.0056	0.6183	0.1078	0.0056	0.3278	0.2411	0.0056	0.0979
	Method 1	0.0500	0.0056	0.6167	0.1122	0.0056	0.3278	0.2444	0.0056	0.0979
	Method 2	0.0511	0.0056	0.6067	0.1111	0.0056	0.3278	0.2411	0.0056	0.0979
	Method 3	0.0511	0.0056	0.6067	0.1133	0.0056	0.3178	0.2422	0.0056	0.0979
	Method 4	0.0511	0.0056	0.6067	0.1167	0.0056	0.3178	0.2511	0.0056	0.0879
1b	Graph A	0.0567	0.0072	0.5783	0.1311	0.0072	0.2775	0.2722	0.0072	0.0748
	Graph B	0.0578	0.0072	0.5683	0.1333	0.0078	0.2792	0.2811	0.0078	0.0757
	Method 1	0.0567	0.0072	0.5783	0.1367	0.0078	0.2787	0.2956	0.0078	0.0752
	Method 2	0.0567	0.0072	0.5783	0.1344	0.0072	0.2770	0.2867	0.0072	0.0740
	Method 3	0.0578	0.0072	0.5773	0.1389	0.0072	0.2750	0.3033	0.0072	0.0740
	Method 4	0.0600	0.0072	0.5673	0.1456	0.0072	0.2750	0.3200	0.0078	0.0643
2a	Graph A	0.0900	0.0183	0.4625	0.2111	0.0183	0.2235	0.4167	0.0183	0.0869
	Graph B	0.0911	0.0178	0.4575	0.2144	0.0183	0.2018	0.4244	0.0194	0.0904
	Method 1	0.0956	0.0183	0.4508	0.2278	0.0189	0.1910	0.4622	0.0228	0.0976
	Method 2	0.0933	0.0183	0.4608	0.2144	0.0178	0.2143	0.4400	0.0183	0.0836
	Method 3	0.1022	0.0178	0.4454	0.2233	0.0178	0.1900	0.4667	0.0194	0.0754
	Method 4	0.1078	0.0183	0.4425	0.2511	0.0183	0.1756	0.4900	0.0194	0.0615
2b	Graph A	0.0922	0.0183	0.4525	0.2133	0.0189	0.2185	0.4400	0.0206	0.0889
	Graph B	0.0944	0.0178	0.4508	0.2167	0.0178	0.1943	0.4522	0.0194	0.0870
	Method 1	0.1033	0.0189	0.4435	0.2344	0.0200	0.1782	0.4822	0.0261	0.1041
	Method 2	0.1033	0.0178	0.4404	0.2167	0.0178	0.2177	0.4633	0.0183	0.0832
	Method 3	0.1100	0.0189	0.4392	0.2511	0.0183	0.1781	0.4889	0.0200	0.0649
	Method 4	0.1200	0.0189	0.4292	0.2933	0.0194	0.1687	0.5378	0.0222	0.0661
3a	Graph A	0.1356	0.0350	0.4195	0.2811	0.0361	0.2200	0.5022	0.0361	0.1202
	Graph B	0.1344	0.0356	0.4249	0.2856	0.0361	0.2177	0.5067	0.0383	0.1245
	Method 1	0.1400	0.0361	0.4205	0.2978	0.0378	0.2209	0.5222	0.0394	0.1225
	Method 2	0.1389	0.0356	0.4230	0.2867	0.0356	0.2155	0.5089	0.0361	0.1188
	Method 3	0.1422	0.0361	0.4224	0.2967	0.0356	0.2094	0.5189	0.0367	0.1184
	Method 4	0.1489	0.0356	0.4143	0.3156	0.0356	0.2010	0.5322	0.0367	0.1160
3b	Graph A	0.1322	0.0317	0.4226	0.2911	0.0322	0.1930	0.5211	0.0322	0.1045
	Graph B	0.1344	0.0317	0.4253	0.3000	0.0322	0.2042	0.5278	0.0339	0.1073
	Method 1	0.1522	0.0333	0.4028	0.3311	0.0361	0.2048	0.5656	0.0394	0.1131
	Method 2	0.1400	0.0311	0.4208	0.3100	0.0317	0.1964	0.5322	0.0317	0.1004
	Method 3	0.1489	0.0311	0.4108	0.3456	0.0317	0.1739	0.5778	0.0317	0.0944
	Method 4	0.1622	0.0322	0.4003	0.3933	0.0333	0.1675	0.6267	0.0339	0.0906

(continued)

Table 3 (continued)

Group	Method	Weak association			Median association			Strong association		
		TPR	FPR	FDR	TPR	FPR	FDR	TPR	FPR	FDR
4a	Graph A	0.2089	0.0689	0.4314	0.3867	0.0700	0.2772	0.5989	0.0706	0.1828
	Graph B	0.2100	0.0678	0.4244	0.3867	0.0672	0.2681	0.5967	0.0667	0.1744
	Method 1	0.2133	0.0694	0.4206	0.3911	0.0694	0.2719	0.6011	0.0700	0.1802
	Method 2	0.2089	0.0672	0.4288	0.3867	0.0672	0.2706	0.6011	0.0678	0.1770
	Method 3	0.2133	0.0672	0.4238	0.3967	0.0678	0.2565	0.6044	0.0672	0.1748
	Method 4	0.2144	0.0683	0.4248	0.4056	0.0683	0.2549	0.6056	0.0678	0.1753
4b	Graph A	0.1789	0.0456	0.3974	0.3844	0.0467	0.2107	0.5989	0.0478	0.1287
	Graph B	0.1822	0.0489	0.4124	0.3889	0.0506	0.2161	0.6167	0.0528	0.1332
	Method 1	0.2022	0.0528	0.3817	0.4389	0.0544	0.1998	0.6644	0.0594	0.1413
	Method 2	0.1800	0.0467	0.4134	0.4056	0.0478	0.1956	0.6256	0.0489	0.1243
	Method 3	0.2011	0.0494	0.3961	0.4444	0.0506	0.1937	0.6589	0.0528	0.1251
	Method 4	0.2522	0.0506	0.3425	0.4989	0.0533	0.1758	0.7133	0.0544	0.1219

Table 4 Details of lung cancer datasets

Dataset name	Number of controls	Number of cases
CL	17	65
Moff	27	52
NCIU133A	18	86
NCILungU133A	44	131

6 Lung Cancer Data

We used four mRNA microarray datasets of lung adenocarcinoma [43] to evaluate the performances of the proposed methods. Data were pre-processed and patients were grouped to two categories, labeled as cases and controls, according to their survival times. For details of the data processing, please see [11]. Each of the datasets has 12,992 genes. Table 4 contains information of the data example. Two-sample t-tests were used to obtain p-values for all genes in all four datasets.

We used 59 lung cancer genes [11] as true positive genes in our analysis. However, this set has a much smaller size than the number of genes in the study. To find additional "positive" genes, CL, Moff, and NCIU133A were used as discovery datasets. For every gene, we used Fisher's Method to combine the three p-values from the three discovery sets to obtain an overall p-value. Then, we defined new positive genes by controlling the FDR under 0.15 using the Benjamini–Hochberg procedure [4]. As a result, among the 12,992 genes, a total of 1044 (or 8.0%) are positive genes in the end.

We extracted 528 biological pathways from KEGG (http://www.kegg.jp), GennMapp (http://genmapp.org), and BioCarta (http://www.biocarta.com) that contained 3735 unique genes, among which 301 ones (or 8.1%) are in the positive set. We found that 379 pathways had at least one lung cancer-associated gene. Finally, we

Table 5 Information of 3 biological pathways

Pathway name (short name)	Number of genes	Number of true positive genes under an FDR cutoff of 0.15
GM human integrin-mediated cell adhesion (adhesion)	118	10
GM human regulation of actin cytoskeleton (regulation)	206	12
GM human signaling of hepatocyte growth factor receptor (HGFR)	120	11

Table 6 AUC of single- and combined-pathway analyses for 14 positive genes under FDR cutoff of 0.15

	Group							
Method	1a	1b	2a	2b	3a	3b	4a	4b
Adhesion	0.5851	0.6425	0.6437	0.6586	0.6400	0.6550	0.6465	0.6693
Regulation	0.5743	0.6516	0.6738	0.7042	0.6448	0.6432	0.5945	0.5438
PGFR	0.5762	0.6650	0.6460	0.6773	0.6752	0.6821	0.6590	0.6894
Method 1	0.5768	0.6511	0.6555	0.6899	0.6641	0.6714	0.5745	0.5768
Method 2	0.5869	0.6722	0.7041	0.7175	0.7050	0.7057	0.7162	0.6920
Method 3	0.5859	0.6749	0.7033	0.7058	0.7068	0.7031	0.6952	0.6586
Method 4	0.5879	0.6835	0.7047	0.6971	0.6929	0.6913	0.6701	0.6362

chose three GennMapp (GM) pathways that were enriched with true positive genes. Table 5 shows details about these three pathways.

The combined pathway had 256 distinct genes, 14 of which were associated with lung cancer (Table 5). We used p-values of dataset NCILungU133A as our test data. However, this test set was missing 18 of the 256 genes in the combined gene pathway. After these 18 genes were removed, the combined pathway has 238 genes including 14 positive genes. To conduct a fair comparison between using a single pathway versus using multiple ones, the same genes had to appear in both the single pathway and the combined one. Therefore, the genes that were in the combined pathway but not in the single pathway were added as singletons to each single pathway.

The same 8 sets of prior parameters (Table 1) were used in our analysis. Use Gibbs sampling to draw random samples from the posterior distribution. Restart the Gibbs sampler 50 times and iterate 1000 times (the first 300 were burn-ins) for each restart. Calculate AUC for all methods. The AUC based on the p-values is 0.5663. Table 6 and Fig. 10 show the values of the AUC for 3 single pathways and 4 proposed methods on the combined pathway.

The black horizontal line in Fig. 10 represents the value of AUC obtained from p-values alone. In general, incorporating pathways, using either a single pathway or the combined one, outperformed the one using p-value only. Comparing the performance of using the combined pathway to one single pathway, the AUCs of Methods 2, 3, and 4 are higher than using a single pathway in all settings. However,

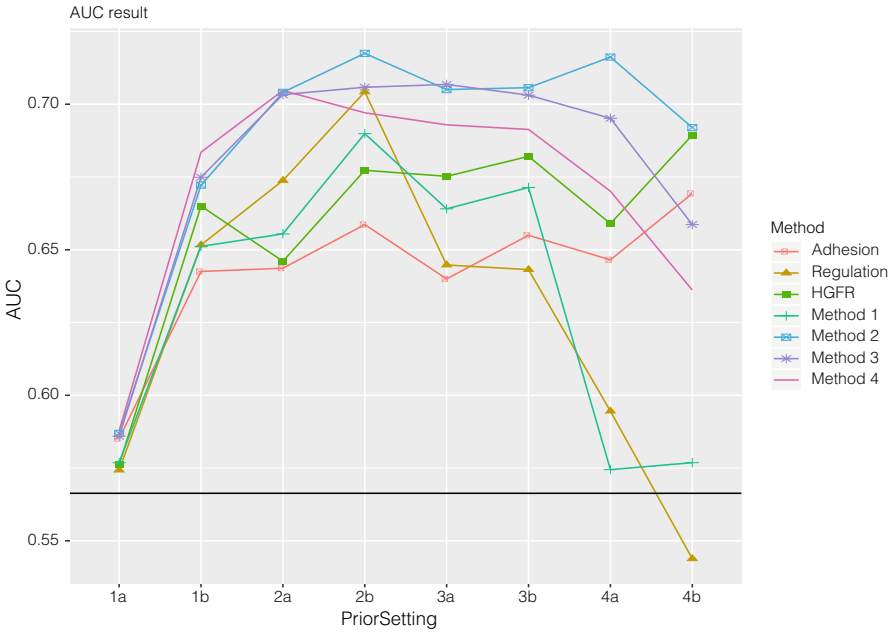


Fig. 10 AUC of single- and combined-pathway analyses with 3 GennMapp pathways to identify 14 positive genes (under an FDR cutoff of 0.15). Adhesion, Regulation, and HGFR refer to the 3 single-pathway analyses. Methods 1–4 are the proposed methods to combine the 3 pathways

Table 7 AUC of single- and combined-pathway analyses for 8 positive genes under FDR cutoff of 0.10

	Group							
Method	1a	1b	2a	2b	3a	3b	4a	4b
Adhesion	0.6285	0.6970	0.7128	0.6967	0.7122	0.7174	0.7011	0.7476
Regulation	0.6114	0.6739	0.6728	0.6938	0.6516	0.6527	0.6177	0.5747
PGFR	0.6084	0.6788	0.6549	0.6755	0.7057	0.6989	0.6736	0.7698
Method 1	0.6120	0.6663	0.6413	0.6905	0.7076	0.6935	0.6133	0.6087
Method 2	0.6302	0.7326	0.7302	0.7258	0.7505	0.7269	0.7644	0.7399
Method 3	0.6274	0.7457	0.7337	0.7122	0.7571	0.7370	0.7454	0.7166
Method 4	0.6293	0.7579	0.7446	0.7253	0.7435	0.7217	0.7255	0.6859

Method 1 does not work well. One possible reason is that there are more shared edges between two nodes in the combined network, but Method 1 only gives weight to the neighbor nodes and does not consider the edges between the nodes.

To examine the impact of selecting positive genes, we further chose a smaller set of positive genes by controlling the FDR at 0.10 instead of 0.15, and it produced 660 positive genes. In the 3 pathways considered above, 8 genes are in the positive set, and the AUC based on the p-values is 0.6571. We reconducted the analysis, and Table 7 reports values of the AUC for the pathway analysis. With the exception of

prior parameter 1a, in general, the pathway-based analyses are better than using p-value alone, and combining 3 pathways by Methods 2, 3, and 4 outperforms the single-pathway methods. However, Method 1 does not work well with prior parameters 4a and 4b.

7 Discussion

We propose to integrate multiple biological pathways to identify disease-related genes. The proposed methods extend the approach of Chen et al. [10] from a single biological pathway to multiple biological pathways. The proposed methods are different from pathway-based approaches that do not take topological structure into account. Simulation studies show that the proposed methods can outperform the methods that only use a single biological pathway. Also, the performances of proposed methods are evaluated with the lung cancer data.

There are some challenges that have influences on the performance of topological-based approaches. First, the inaccuracy and incompleteness of biological pathways can lead to the loss of statistical power. In the biological pathways, some genes interact with others through chemical compounds. However, biological pathways extracted from online databases will lose such gene–compound interactions if we focus on genes. For example, gene NOD2 has been identified significantly associated with Crohn' s disease [16]. However, NOD2 indirectly interacts with other genes in the Inflammatory Bowel Disease pathway, and it becomes an isolated gene in the pathway extracted from KEGG. When we apply the proposed approaches, NOD2 has been removed because of the loss of compound mediated interactions. A number of isolated genes can lead to the loss of information about gene–gene interactions. Moreover, it can reduce the statistical power of topological-based approaches. Second, the inconsistency of biological pathways from different data bases [14, 31] can lead to inconsistent conclusions. For instance, gene ontology (GO) [1] defines different pathways for apoptosis in different cell types. Alternatively, KEGG only defines a single pathway for apoptosis. The different definitions of biological pathways in different data bases can affect the results of the approaches. Third, the choices of biological pathways have an influence on the results. When we choose biological pathways that are used to generate combined pathway, we choose the ones that are related to the disease. In general, opinions from experts and external resources are required. Fourth, as the size of biological pathway increases, the computational task will become more intensive.

There is a limitation that may affect the performance of the proposed approaches. The prior setting varies with the sizes and structures of biological pathways. For estimating the hyper-parameters, in the Supplementary Text S2 of [10], the authors described a conditional empirical Bayes approach, which can be readily applied to this chapter. For the future work, distributions may be considered to the hyper-parameters to account for the variability of these parameters.

Acknowledgments This work was partially supported by the National Institutes of Health [grant R15GM131390 to X.W.].

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nat. Genet. **25**(1), 25–29 (2000)
2. Bair, E., Hastie, T., Paul, D., Tibshirani, R.: Prediction by supervised principal components. J. Am. Stat. Assoc. **101**(473), 119–137 (2006). http://www.jstor.org/stable/30047444
3. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nat. Rev. Genet. **12**(1), 56–68 (2011)
4. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. **57**(1), 289–300 (1995). http://www.jstor.org/stable/2346101
5. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Methodol. **36**(2), 192–236 (1974)
6. Bokanizad, B., Tagett, R., Ansari, S., Helmi, B.H., Draghici, S.: SPATIAL: A System-level PAThway Impact AnaLysis approach. Nucl. Acids Res. **44**(11), 5034–5044 (2016)
7. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al.: The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucl. Acids Res. **47**(D1), D1005–D1012 (2019)
8. Bush, W.S., Moore, J.H.: Genome-wide association studies. PLoS Comput. Biol. **8**(12), e1002822 (2012)
9. Casella, G., George, E.I.: Explaining the Gibbs sampler. Am. Stat. **46**(3), 167–174 (1992)
10. Chen, M., Cho, J., Zhao, H.: Incorporating biological pathways via a Markov random field model in genome-wide association studies. PLoS Genet. **7**(4), 1–13 (2011). https://doi.org/10.1371/journal.pgen.1001353
11. Chen, M., Zang, M., Wang, X., Xiao, G.: A powerful Bayesian meta-analysis method to integrate multiple gene set enrichment studies. Bioinformatics (Oxford, England) **29**, 862–869 (2013). https://doi.org/10.1093/bioinformatics/btt068
12. Chen, X., Wang, L., Hu, B., Guo, M., Barnard, J., Zhu, X.: Pathway-based analysis for genome-wide association studies using supervised principal components. Genet. Epidemiol. **34**(7), 716–724 (2010)
13. Cookson, W., Liang, L., Abecasis, G., Moffatt, M., Lathrop, M.: Mapping complex disease traits with global gene expression. Nat. Rev. Genet. **10**(3), 184–194 (2009)
14. Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., et al.: Pathway and network analysis of cancer genomes. Nat. Methods **12**(7), 615 (2015)
15. Dutta, B., Wallqvist, A., Reifman, J.: Pathnet: a tool for pathway analysis using topological information. Source Code Biol. Med. **7**(1), 1 (2012)
16. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al.: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. **42**(12), 1118–1125 (2010)
17. Freytag, S., Manitz, J., Schlather, M., Kneib, T., Amos, C.I., Risch, A., Chang-Claude, J., Heinrich, J., Bickeböller, H.: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum. Hered. **76**(2), 64–75 (2014)

18. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A.: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. **106**(23), 9362–9367 (2009)
19. Hou, J., Acharya, L., Zhu, D., Cheng, J.: An overview of bioinformatics methods for modeling biological pathways in yeast. Brief. Funct. Genomics **15**(2), 95–108 (2016)
20. Hou, L., Chen, M., Zhang, C.K., Cho, J., Zhao, H.: Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. Hum. Mol. Genet. **23**(10), 2780–2790 (2014)
21. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The reactome pathway knowledgebase. Nucl. Acids Res. **48**, D498–D503 (2020). https://doi.org/10.1093/nar/gkz1031
22. Jin, L., Zuo, X.Y., Su, W.Y., Zhao, X.L., Yuan, M.Q., Han, L.Z., Zhao, X., Chen, Y.D., Rao, S.Q.: Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics **12**(5), 210–220 (2014)
23. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: Data, information, knowledge and principle: back to metabolism in KEGG. Nucl. Acids Res. **42**(D1), D199–D205 (2014)
24. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: KEGG as a reference resource for gene and protein annotation. Nucl. Acids Res. **44**(D1), D457–D462 (2016)
25. Krauss, G.: Biochemistry of Signal Transduction and Regulation. Wiley, London (2006)
26. Lin, Z., Li, M., Sestan, N., Zhao, H.: A markov random field-based approach for joint estimation of differentially expressed genes in mouse transcriptome data. Stat. Appl. Genet. Mol. Biol. **15**(2), 139–150 (2016)
27. Liu, J., Peissig, P., Zhang, C., Burnside, E., McCarty, C., Page, D.: Graphical-model based multiple testing under dependence, with applications to genome-wide association studies. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, vol. 2012, p. 511. NIH Public Access (2012)
28. Liu, L., Lei, J., Roeder, K., et al.: Network assisted analysis to reveal the genetic basis of autism. Ann. Appl. Stat. **9**(3), 1571–1600 (2015)
29. Loscalzo, J., Kohane, I., Barabasi, A.L.: Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol. Syst. Biol. **3**(1), 124 (2007)
30. Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C.I., Xiong, M.: Genome-wide gene and pathway analysis. Eur. J. Hum. Genet. **18**(9), 1045–1053 (2010)
31. Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C., Draghici, S.: Methods and approaches in the topology-based analysis of biological pathways. Front. Physiol. **4**(278), 1–22 (2013)
32. Mokry, M., Middendorp, S., Wiegerinck, C.L., Witte, M., Teunissen, H., Meddens, C.A., Cuppen, E., Clevers, H., Nieuwenhuis, E.E.: Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. Gastroenterology **146**(4), 1040–1047 (2014)
33. Mourad, R., Sinoquet, C., Leray, P.: Probabilistic graphical models for genetic association studies. Brief. Bioinform. **13**(1), 20–33 (2012)
34. Nica, A.C., Dermitzakis, E.T.: Using gene expression to investigate the genetic basis of complex disorders. Hum. Mol. Genet. **17**(R2), R129–R134 (2008)
35. Pan, W.: Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet. Epidemiol. **35**(4), 211–216 (2011). https://doi.org/10.1002/gepi.20567
36. Pan, W., Kim, J., Zhang, Y., Shen, X., Wei, P.: A powerful and adaptive association test for rare variants. Genetics **197**(4), 1081–95 (2014). https://doi.org/10.1534/genetics.114.165035
37. Pan, W., Kwak, I.Y., Wei, P.: A powerful pathway-based adaptive test for genetic association with common or rare variants. Am. J. Hum. Genet. **97**(1), 86–98 (2015). https://doi.org/10.1016/j.ajhg.2015.05.018

38. Pavlopoulos, G.A., Secrier, M., Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., Bagos, P.G.: Using graph theory to analyze biological networks. BioData Mining **4**(1), 1 (2011)
39. Rapin, N., Bagger, F.O., Jendholm, J., Mora-Jensen, H., Krogh, A., Kohlmann, A., Thiede, C., Borregaard, N., Bullinger, L., Winther, O., et al.: Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. Blood **123**(6), 894–904 (2014)
40. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S., Collins, A.L., Crowley, J.J., Fromer, M., et al.: Genome-wide association analysis identifies 14 new risk loci for schizophrenia. Nat Genet. **45**(10), 1150–1159 (2013)
41. Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., Mistry, H., Mosier, L., Dlin, J., Wen, Q., O'Callaghan, C., Li, W., Elder, G., Smith, P.T., Dallago, C., Cerami, E., Gross, B., Dogrusoz, U., Demir, E., Bader, G.D., Sander, C.: Pathway commons 2019 update: integration, analysis and exploration of pathway data. Nucl. Acids Res. **48**, D489–D497 (2020). https://doi.org/10.1093/nar/gkz946
42. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**(11), 2498–2504 (2003)
43. Shedden, K., Taylor, J.M.G., Enkemann, S.A., Tsao, M.S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., Misek, D.E., Chang, A.C., Zhu, C.Q., Strumpf, D., Hanash, S., Shepherd, F.A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Sharma, A., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., Conley, B., Seshan, V.E., Meyerson, M., Kuick, R., Dobbin, K.K., Lively, T., Jacobson, J.W., Beer, D.G.: Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat. Med. **14**, 822–827 (2008). https://doi.org/10.1038/nm.1790
44. Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al.: Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucl. Acids Res. **46**(D1), D661–D667 (2018)
45. Song, G.G., Lee, Y.H.: Pathway analysis of genome-wide association study on asthma. Hum. Immunol. **74**(2), 256–260 (2013)
46. Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P., Romero, R.: A novel signaling pathway impact analysis. Bioinformatics **25**(1), 75–82 (2009)
47. Wei, P., Pan, W.: Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. Ann. Appl. Stat. **6**(1), 334 (2012)
48. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., Lin, X.: Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. **86**(6), 929–942 (2010)
49. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X.: Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. **89**(1), 82–93 (2011). https://doi.org/10.1016/j.ajhg.2011.05.029
50. Zalkin, H., DAGLEY, S., Nicholson, D.E.: An Introduction to Metabolic Pathways. Wiley, London (1971)
51. Zhi, W., Minturn, J., Rappaport, E., Brodeur, G., Li, H.: Network-based analysis of multivariate gene expression data. In: Statistical Methods for Microarray Data Analysis: Methods and Protocols, pp. 121–139 (2013)