Hindawi
BioMed Research International
Volume 2019, Article ID 2497509, 12 pages
https://doi.org/10.1155/2019/2497509

Review Article

Incorporating Pathway Information into Feature Selection towards Better Performed Gene Signatures

Suyan Tian,[1] **Chi Wang**,[2] **and Bing Wang**[3]

[1]*Division of Clinical Research, The First Hospital of Jilin University, 71 Xinmin Street, Changchun, Jilin 130021, China*
[2]*Department of Biostatistics, Markey Cancer Center, The University of Kentucky, 800 Rose St., Lexington, KY 40536, USA*
[3]*School of Life Science, Jilin University, 2699 Qianjin Street, Changchun, Jilin 130012, China*

Correspondence should be addressed to Suyan Tian; windytian@hotmail.com and Chi Wang; chi.wang@uky.edu

Received 23 July 2018; Accepted 7 March 2019; Published 3 April 2019

Academic Editor: Paul Harrison

To analyze gene expression data with sophisticated grouping structures and to extract hidden patterns from such data, feature selection is of critical importance. It is well known that genes do not function in isolation but rather work together within various metabolic, regulatory, and signaling pathways. If the biological knowledge contained within these pathways is taken into account, the resulting method is a pathway-based algorithm. Studies have demonstrated that a pathway-based method usually outperforms its gene-based counterpart in which no biological knowledge is considered. In this article, a pathway-based feature selection is firstly divided into three major categories, namely, pathway-level selection, bilevel selection, and pathway-guided gene selection. With bilevel selection methods being regarded as a special case of pathway-guided gene selection process, we discuss pathway-guided gene selection methods in detail and the importance of penalization in such methods. Last, we point out the potential utilizations of pathway-guided gene selection in one active research avenue, namely, to analyze longitudinal gene expression data. We believe this article provides valuable insights for computational biologists and biostatisticians so that they can make biology more computable.

1. Introduction

Data obtained from the high-throughput technologies such as microarrays or RNA-sequencing (RNA-seq) is a recurring theme in many fields such as computational biology and bioinformatics. Given these advanced technologies are expensive, the number of observations/subjects is usually small, i.e., on the scales of several to hundreds. Another special characteristic of the high-throughput technologies is that they can measure thousands of variables/features simultaneously. As far as the statistical modeling is considered, a classic regression model becomes nonidentifiable when all measured variables are used as predictors for such a data set; let alone one may also be interested in exploring the nonlinear association at higher orders or the interactions among these variables. To deal with the data in which the number of variables is extremely larger than the number of samples, the implementation of a feature selection process that identifies a subset of genes with the optimal predictive performance [1] is in demand.

Feature selection has outstanding merits. Especially, the resulting subset of genes speeds up the learning process, improves predictive accuracy, and leads to a better biological implication. The classic feature selection, we call it "gene-based feature selection" to avoid ambiguity in this article, is stratified into three subtypes, say, filter, embedded, and wrapper methods [1, 2]. These three categories have their own unique characteristics. For instance, a filter method usually screens individual features one by one according to their relevancy level with the outcome of interest [1]. The feature selection of an embedded method is usually realized by using a penalized regression model such as the Least Absolute Shrinkage and Selection Operator (LASSO) model [3]. Such a method can simultaneously select relevant features and estimate those coefficients (the effect size of those features) in the final model; in addition to that it consumes less computing time than a wrapper method.

Nevertheless, the gene subset/list selected by a gene-based feature selection algorithm has several drawbacks. First,

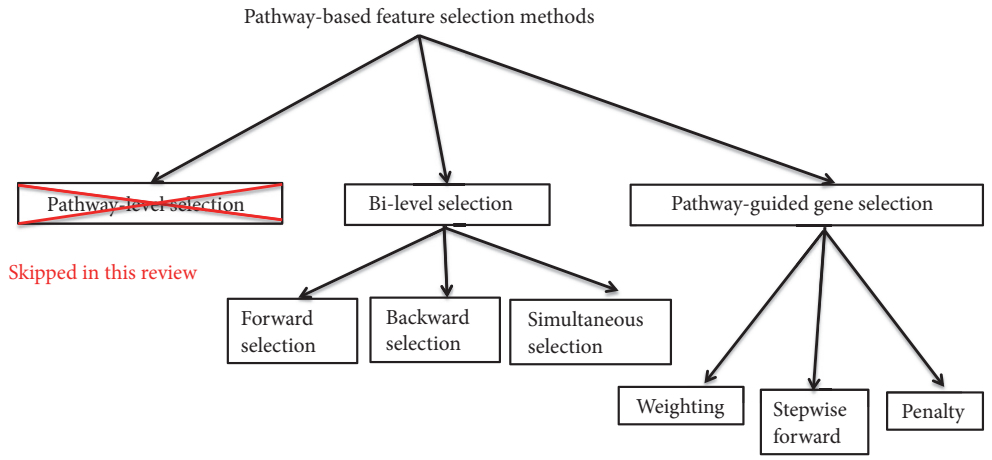


Figure 1: Major ramifications of pathway-based feature selection methods.

the predictive performance on new independent samples is unsatisfactory; the overfitting phenomenon is always apparent. Second, the gene lists trained from different data sets barely overlap. Reproducibility or stability of the final models (with different data, the same method gives different gene lists with few or no overlaps) is low, leading to a generalization of the resulting gene list impossible. Last, most of these methods use the difference of gene expression level between different phenotypes as a critical criterion to select the genes associated with the outcome. However, differentially expressed genes (DEGs) are not necessarily to be true driver genes. Ignoring biological information may result in a meaningless biological implication for the resultant gene list.

Furthermore, it is well known that genes do not function in isolation but rather work together within various metabolic, regulatory, and signaling pathways. The interdependencies among genes are often represented as a collection of pathways/gene sets in which potential coregulated or coexpressed genes are grouped together. In this review, the terms "pathway", "network", and "gene set" have same implication/meaning and are exchangeable to one another.

The biological information contained within pathways can be utilized to impose additional constraints on the prediction tasks, forcing training methods to select more scientifically meaningful genes rather than those statistically significant genes (such as genes more differentially expressed between two phenotypes). A feature selection process that incorporates pathway knowledge by one means or another is referred to as pathway-based feature selection herein, which has currently grown into a hot topic in computational biology and bioinformatics.

So far, to the best of our knowledge, no survey on pathway-based feature selection methods in the literature has been given yet. The objective of this article is to provide a selective review on such methods.

2. Pathway-Based Feature Selection Methods

Based on to what a feature refers, pathway-based feature selection methods may be classified into three categories; see Figure 1. The first category contains pathway analysis methods such as [4, 5] in which a feature corresponds to a pathway, with the objectives of selecting the whole pathways associated with the phenotypes of interest. Since the methods in this category have been reviewed by many researchers previously and several well-known algorithms have been compared exclusively using simulations and real-world data [6–15] and our attention is mainly focused on the selection of individual genes related to the phenotypes of interest, we skip this topic in the article.

The second category considers a bilevel selection process, which identifies not only relevant pathways but also important genes that contribute critically to the significance of identified relevant pathways. The bilevel selection methods can be further divided into three major categories, forward selection, backward selection, and simultaneous selection [16]. In a forward selection process such as [10], the selection of relevant gene sets is carried out firstly and then followed by the selection of relevant individual genes. In contrast, the selection steps in a backward selection process such as [17] take the reversed order. Last, a simultaneous selection process such as [18] performs the selection of significant gene sets and the selection of important genes at the same time, as its name implies. The simultaneous selections of gene sets and genes are usually accomplished with the aids of a penalized model where a penalty term imposing some restrictions on the β coefficients that represent the association magnitude with the outcome is added to the objective function.

In the last category, a feature corresponds to an individual gene. The methods of this type incorporate the pathway knowledge as a priori to facilitate the selection of relevant genes, aiming to improve the resulting gene list's predictive ability and/or reproducibility. Although we had intended to reserve the syntax of "pathway-based feature selection" for this specific subfield, we frame a specific term "a pathway-guided gene selection" for it instead to avoid confusions. Given our attention is focused on the methods capable of selecting important individual genes [16], the bilevel selection algorithms, e.g., [10, 19], may be loosely classified into the pathway-guided gene selection category.

3. Pathway-Guided Gene Selection Methods

3.1. Three Major Categories. In our previous study [20], we stratified a pathway-guided gene selection method into three classes on the basis of which piece of pathway information was incorporated and how such information was incorporated, namely, weighting, stepwise forward, and penalty. In the following subsections, a detailed description of and discussion on these three categories are given.

3.2. Stepwise Forward. The stepwise forward methods usually rank all genes according to a specific discriminative score. Then the methods start from the most significant gene and evaluate the performance of the resulting gene subset based on some predetermined metric. The step iterates until no further gain upon this performance statistic can be obtained. A bilevel selection method, the significance analysis of microarray gene set reduction (SAMGSR) algorithm [10], can be put into this category. This method consists of two steps. Its first step is essentially an extension of the significance analysis of microarray (SAM) method [21] to all genes inside a gene set, and a new statistic called SAMGS [4] which is the square sum of SAM statistics for all genes inside a specific gene set is generated. The significance level of a gene set is determined using permutation tests. Obviously, this step carries out the selection of significant pathways firstly so that the SAMGSR method belongs to the forward bilevel selection category. In the second step, a subset of important genes is extracted from each significant pathway identified by the first step on the basis on the magnitudes of individual genes' SAM statistics. The realization of this extraction is by the means of stepwise forward. Specifically, the genes inside each significant pathway are ordered decreasingly based on the magnitude of their SAM statistics. Then the reduction step gradually partitions the entire gene set into two subsets: the reduced subset that includes the first k genes and the residual subset including the remaining genes for $k = 1,\ldots, |j|$, where $|j|$ is the size of gene set j. At each partition, the significance level of the reduced subset is evaluated using the p-value of SAMGS statistic for its corresponding residual subset. The iteration stops until this p-value is larger than a predetermined threshold for the first time.

Another typical example of a stepwise forward method is the algorithm proposed by Chuang et al. [22]. This method starts from a seed gene and identifies a gene list by gradually adding the neighboring gene that provides the highest mutual information between the average of expression values for all included genes and the outcome. In this example, network topology information that records how genes are connected instead of the grouping membership information is taken into consideration.

Two big drawbacks of a stepwise forward method are as follows: (1) the methods may fail to identify those 'driving' genes with subtle changes because the inclusion of a gene depends largely on its expression values or expression differences among different phenotypes; (2) the selection process of important genes is usually separated from the final model construction.

3.3. Weighting. The weighting methods construct a pathway knowledge-based weight that reflects how important a gene is inside the gene-to-gene interaction network for each gene and then balance between the weight and its gene expression values to determine the significance level of the specific gene. For example, the reweighted recursive feature elimination (RRFE) method [23] uses the GeneRank algorithm [24] to alter the ranking criterion of the support vector machine recursive feature elimination (SVM-RFE) algorithm, and then identifies a subset with the best discriminative power. More specifically, the resulting GeneRanks are used as weights and combined with the coefficients of SVM to increase the chance of a gene with more directly connected neighbors being selected.

In the RRFE method, the weights are combined with the statistics (i.e., the coefficients in a SVM model). An alternative strategy of weighting is to combine the weights directly with gene expression values to generate weighted gene expression values and then implement a gene-based feature selection method such as LASSO to identify relevant genes. An example of this category is [25], in which the weighted expression profiles were used to classify two major subtypes of non-small-cell lung cancer. Overall, the weighting methods are the least implemented in the literature, compared to the methods in other two categories. This may be due to that the constructed weights are subject to biases and errors, which might lead to inferiority of the resulting gene lists.

3.4. Penalty. In a penalty model, an extra penalty term that records pathway information is combined with an objective function such as the log likelihood function to generate the final objective function. The identification of relevant genes is realized by the means of finding the best subset of genes that optimize this function. To name several penalty methods, Zhu et al. [26] combined the network-constrained penalty term given by [27] with a SVM model and proposed the network-based SVM method to discriminate two different phenotypes. Similarly, Chen et al. [28] also combined the network-constrained penalty term with a SVM model and proposed the netSVM method for the purpose of classification. More recently, Sokolov et al. [29] generalized the elastic net penalty term to incorporate pathway knowledge and then combined the proposed penalty term with an objective function to select relevant genes. The proposed term is referred to as the generalized elastic net (gelnet) function, and it includes the elastic net as a special case. The big disadvantage of a penalty method is that its computing burden is moderate or even heavy. Three separate figures (Figures 2–4) were made to elucidate these three major types of pathway-guided gene selection methods in detail. A review of typical examples in each category is given in Table 1.

3.5. Penalty Function. Given the fact that penalization plays a critical role in both the pathway-guided gene selection and in bilevel selection methods, we discuss the commonly used penalty terms in both methods in the following sections.

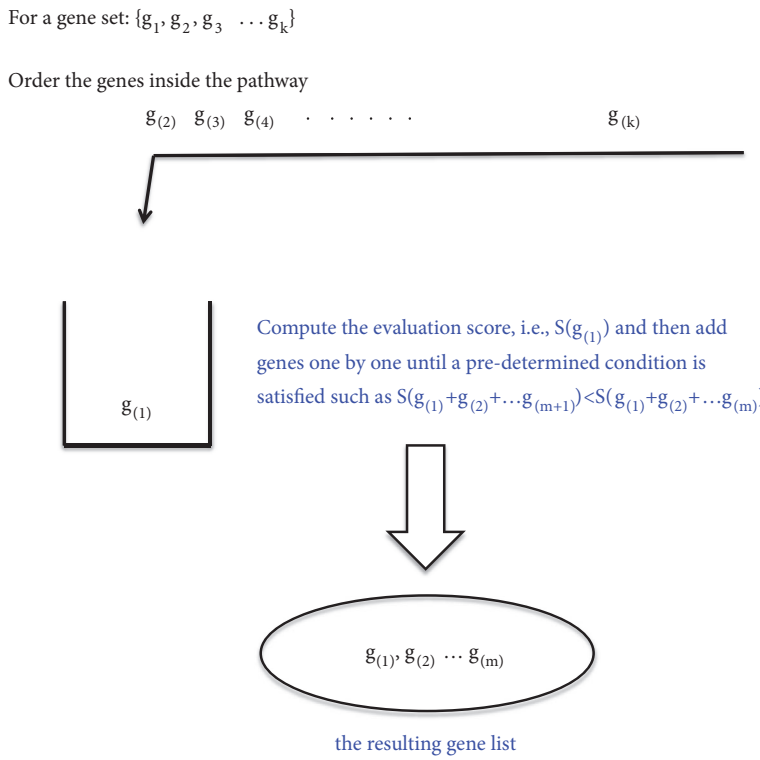


FIGURE 2: Graphical illustration of the stepwise forward methods.

3.5.1. Network-Constrained Penalty. For a penalty method, one well-known network-constrained penalty term was proposed by [27]. It is notated as

$$\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v) \tag{1}$$

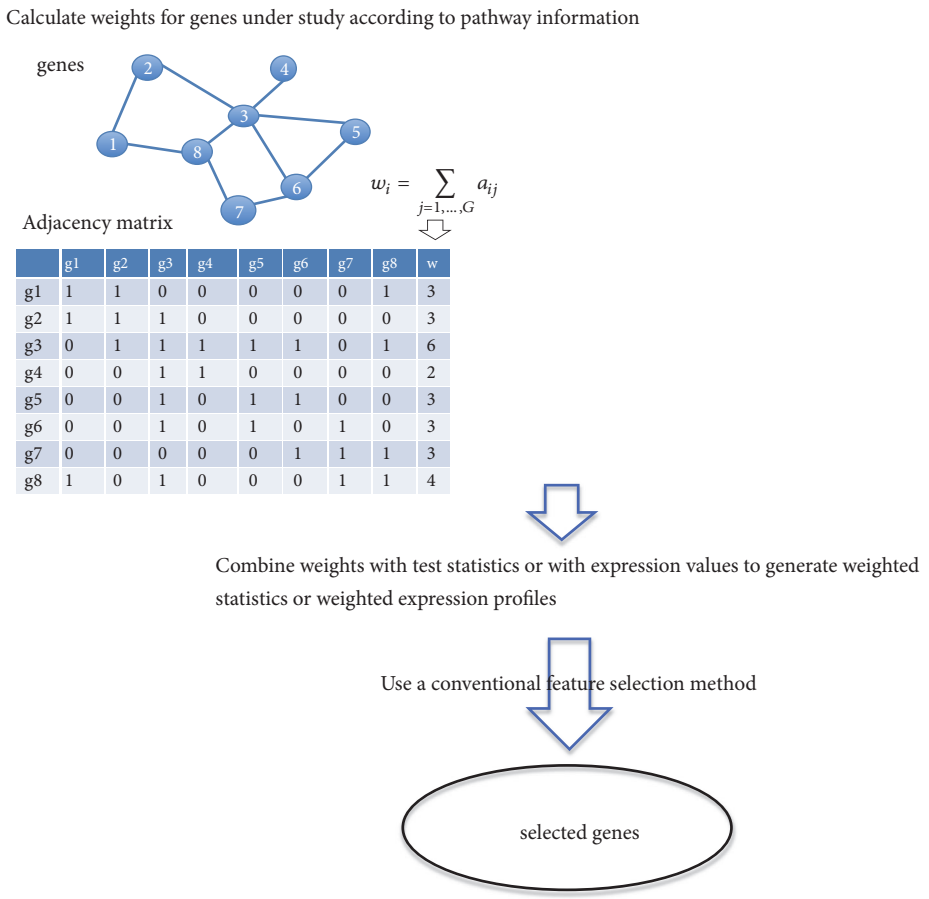
Here w(u,v) denotes the weight of edge (u, v). It usually takes the value of 1 if gene u and gene v are connected, 0 otherwise. The degree of gene u (denoted as d_u) is the sum of edge weights over all vertices connected with u, i.e., $\sum_{u\sim v} w(u,v)$. This term introduces a smooth solution of β coefficients (which represent the association magnitudes and directions of genes with the outcome) on the network via penalizing the weighted sum of squares of the scaled difference of the coefficients between connected genes. Li & Li [27] specifically stated that scaling the β coefficients using their respective degrees of freedom on the network "allows the genes with more connections to have larger coefficients so that small changes of expressions of such genes can lead to large changes in the response". Several studies had adopted and imposed this penalty term on different objective functions. For instance, Chen et al. [28] had imposed this constraint on a support vector machine (SVM) model and developed a new approach called the network-constrained support vector machine (netSVM) method. For a more detailed description on the penalty functions at the pathway level, the work by Pan et al. [30] and Table 2 are referred.

3.5.2. General Penalty Framework for a Bilevel Selection Method. For a bilevel selection process, Breheny & Huang [31] presented a general framework of the penalty functions used, which is

$$f_o\left(\sum_{k=1}^{K_j} f_I\left(\left|\beta_{j,k}\right|\right)\right) \tag{2}$$

where the subscript j, k represents gene k (k=1,2,... K_j, where K_j is the size of gene set j) inside group j (j=1,2,...J, where J is the number of gene sets under consideration). In this formula, an outer penalty function f_o, e.g., the bridge penalty, is applied to a sum of inner penalties f_I, e.g., the LASSO. The outer penalty regularizes the coefficients of all genes within the specific get sets while the inner penalty function penalizes on the coefficients before individual genes. Table 3 summarizes those penalty terms commonly used in a simultaneous bilevel selection process.

After searching in the literature, it is found that pathway-guided gene selection methods have been widely applied in cancer studies. Specifically, a pathway-guided gene selection algorithm may cast some insight on identifying diagnostic gene signatures capable of distinguishing cancer patients from normal controls; different subtypes of a specific cancer; or histologic stages, or identifying prognostic signatures that predict the survival time of cancer patients. By searching in the PubMed using the keywords of feature selection, pathway/network, gene expression, and cancer and then inspecting their relevance, we found roughly 40 articles which utilize pathway-guided gene selection algorithms to study a variety of cancers. Figure 5 provides the statistics of these articles by stratifying them according to the cancer types under study. From this figure, it is observed that the



	g1	g2	g3	g4	g5	g6	g7	g8	w
g1	1	1	0	0	0	0	0	1	3
g2	1	1	1	0	0	0	0	0	3
g3	0	1	1	1	1	1	0	1	6
g4	0	0	1	1	0	0	0	0	2
g5	0	0	1	0	1	1	0	0	3
g6	0	0	1	0	1	0	1	0	3
g7	0	0	0	0	0	1	1	1	3
g8	1	0	1	0	0	0	1	1	4

Figure 3: Graphical illustration of the weighting methods.

most frequently studied cancer types are breast cancer, e.g., the studies by [23, 26, 28, 29, 32–35] and lung cancer, e.g., the studies by [16, 20, 33, 36–40].

Among these studies, the penalty method is the most prevalent method, being followed by the stepwise forward method. This observation provides evidence to support our statement that the strategy of a penalized regression model to select relevant genes has gained increasing attention and the weighting methods have been underutilized compared to the other two categories. Given there are several public repositories such as The Cancer Genome of Atlas (TCGA: https://portal.gdc.cancer.gov/), the Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/), and Array Express [41], we believe more investigation will boom to utilize pathway-guided gene selection methods to study other cancer types and other complex diseases.

4. Pathway Information

4.1. Topology or Grouping Information. As we mentioned in the early section, different algorithms may account for different pathway knowledge. For examples, some algorithms consider pathway topology information (e.g., which genes are connected to which genes) whereas some ignore it. In the methods that omit topology information, genes are grouped into many gene sets and only the group membership of genes is considered. From the perspective of weighting, the methods using grouping information weigh every gene inside a specific pathway equally while the first type of methods may prioritize the genes with high connectivity level. Based on whether topology information is considered, a pathway-guided gene selection method can be divided into either a functional score based method or a topology-based method. In a functional score based method such as [10, 18], only the grouping membership of genes is considered to generate an evaluation score, with an implicit assumption that all genes inside a specific pathway coregulate/cofunction together. In contrast, in a topology-based method such as [28] more structured pathway knowledge rather than grouping information is considered.

4.2. Data-Driven versus Canonical Pathways. Several studies, e.g., [42, 43], have concluded that pathway-guided gene selection does not outperform classic gene-based feature selection methods in terms of predictive accuracy. This inferiority may be explained by the fact that the pathway knowledge retrieved from those canonical pathway databases/knowledge-bases such as the Kyoto Encyclopedia of Gene and Genomes (KEGG) [44], Gene Ontology (GO) [45], and Reactome [46] conveys no or limited meaningful information for a specific dataset or condition/disease. In contrast, the pathways constructed in a "data-driven" way may be more

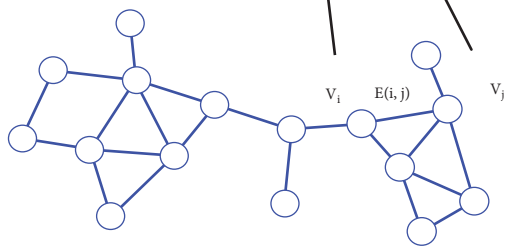
Upon the objective function for different purposes, e.g., when the outcome is a binary variable indicating two distinct groups, the corresponding objective function is as follows

$$\sum_{j=1}^{n}\left(Y_j(\beta_0+\beta X_j)-\log(1+\exp(\beta_0+\beta X_j)\right)$$

Add a gene-based penalty term, e.g., LASSO and a pathway-based penalty term (e.g., the below one)

Pathway-based penalty term $\sum_{(i,j)\in E} w(i,j)\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)^2$

GS: (V, E)



d_i: the degree of freedom for gene i

w(i.j): the connection weight between gene i and j.

e.g., 1 or 0 is used indicate if gene i and j are connected or not, then w(i,j)=1, d_i=3, and d_j=4

Apply an optimization strategy (e.g., the coordinate descent method) to solve the resulting final function, i.e., the objective function + gene-based+ pathway-based penalty terms

Figure 4: Graphical illustration of the penalty methods.

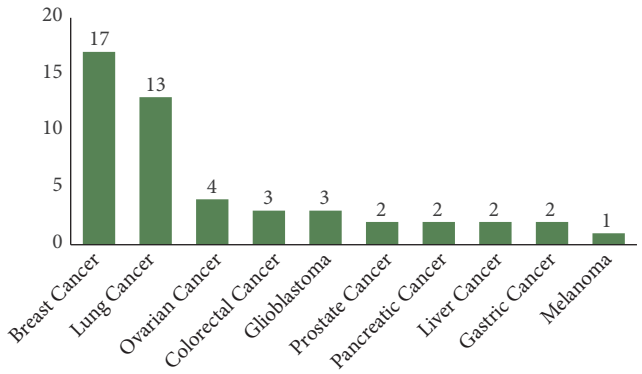


Figure 5: *Statistics for pathway-guided gene selection methods in cancer studies*. A literature search was conducted in the PubMed using keywords of feature selection, gene expression, pathway/network, and cancer. The number of relevant articles stratified by the cancer types under study is given on the top of those bars.

informative for the diseases under investigation and thus be preferred over the canonical pathways. Here, "data-driven" means that specific data of a specific condition/disease are used to build up pathways [47, 48]. The construction for those data-driven pathways is usually accomplished with the aids of a coexpression module detection technique, e.g., the Weighted Gene Coexpression Network Analysis (WGCNA) [49] and Algorithm for the Reconstruction of Accurate Cellular Network (ARACNE) [50]. Also, there exist some elegant algorithms, e.g., [51, 52], that are able to figure out grouping structures and carry out feature selection simultaneously. No matter which strategy it takes, in the "data-driven" pathway construction pathway structure is inferred from data.

On the other hand, data-driven pathways provide no information about causality given they cannot determine genes' positions in the whole network and thus cannot distinguish the regulatory/upstream genes apart from the regulated/downstream genes [61]. In addition, different form the static representations of the biological pathways, say, protein-to-protein interaction, metabolic networks, or signaling networks curated in those canonical databases, these data-driven pathways vary from data to data and thus may be subject to random noises and difficult to be interpretable from a biological point of view [62]. Finally, the resulting models by using data-driven pathways may be subject to overfitting since the build-up of coexpression/coregulation

Table 1: A selective review of pathway-guided gene selection algorithms.

Reference	Brief description of the proposed method and its characteristics	Category
Zhu et al. [26]	The proposed network-based SVM method combines the network-constrained penalty (see equation (1)) with a SVM model to carry out feature selection and classification. It makes SVM models capable of carrying out feature selection; the network-constrained penalty gives heavier weights to genes with more direct neighbors (thus increases the chance of such genes being selected) and encourages a grouping effect. But the method only deals with binary classification and considers immediate neighbors.	Penalty
Chen et al. [28]	The netSVM method also combines the network-constrained penalty (see equation (1)) with a SVM model. Its advantages and disadvantages are similar to the network-based SVM method by Zhu et al [26] (see above)	Penalty
Sokolov et al. [29]	The generalized elastic net penalty function is given and combined with an objective function to select important genes. This is named as the GELnet method. The authors claimed that this penalty function includes many well-known penalty terms and the method is so flexible that it can deal with many outcome types. There is an independent R package (i.e., gelnet) to implement this method, but now this package can only conduct binary classification.	Penalty
Zhang et al. [53]	The Net-Cox method adds a network-constrained penalty term to the corresponding partial likelihood function of a Cox model, aiming to select important prognostic genes The Matlab codes are available online, making the implementation of this method easy. This method only considers direct neighbors.	Penalty
Bandyopadhyay et al. [32]	After ranking genes in a pathway according to their marginal classification power, the proposed BPFS method starts from the gene with the largest power and then adds genes The authors claimed that this method goes beyond the immediate neighbors and considers redundant gene elimination. Also, missing genes in the pathway databases are mapped to the network using a probabilistic technique. However, the method is hard to comprehend, and no codes are available.	Stepwise forward
Lee et al. [33]	In each pathway, the method reorders genes according to their t-scores, and then the subset of genes whose combined expression has optimal discriminative power called CORGs is identified. Only the membership of genes is considered. The method is simple and easy to implement.	Stepwise forward
Razi et al. [34]	The proposed NBCG method starts with a seed gene and traverses the network to find the optimal subset on the basis of Shapley value. The method uses the concept of Shapley value to take into account the collective power of the resulting gene subset. The choice of a seed gene may result in excluding a gene subset with subtle individual effects but significant concordant effect.	Stepwise forward
Wu et al. [54]	The shortest path method (with well-known genes related to the disease under study, i.e., gastric cancer as seeds) is used to mine candidate genes and the combination of random forest +incremental feature selection is used to obtain the optimal subset. The proposed method considers topology information of a network. The use of a wrapper method (RF+IFS) and permutation tests may slow the method down.	Stepwise forward[1]
Tian et al. [20]	The weighted-SAMGSR method extends the SAMGSR algorithm by weighing SAMGS statistics according to genes' connectivity levels in the network. The method considers both the membership information and the connectivity level, and can handle two-class and multiple-class classification. The R-codes are available in the supplementary material. Computing time is a big concern since permutation tests are needed to calculate p-values of test statistics.	A hybrid of weighting and stepwise forward

Table 1: Continued.

Reference	Brief description of the proposed method and its characteristics	Category
Johannes et al. [23]	The RRFE method uses the GeneRank algorithm to alter the ranking criterion of the SVM-RFE algorithm and selects a subset with the best discriminative power. Weighing the coefficients of SVM models with their GeneRanks to increase the probability of a gene with more connected genes being selected, an independent R package (i.e., pathClass) is provided to implement this method. The method only considers how many direct neighbors a gene has and ignores topology information completely.	Weighting
Chan et al. [39]	The wgSVM-SCAD method weighs the expression values of genes in a pathway according to their t-values and then uses a penalized SVM model (with SCAD penalty) to identify relevant genes. The proposed method only considers membership information and the weights are only based on the relevance score (i.e., t-values) instead of pathway information.	Weighting
Tian et al. [16]	Using sign averages of all genes inside a gene set to represent corresponding gene set, the proposed methods (i.e., one forward bi-level selection method and one backward bi-level selection method) filter out insignificant gene sets and insignificant genes in a specific order. The sign average metric provides a better representation of a gene set than mean, median and the first PC. The proposed methods only consider membership information.	Bi-level selection
Lim and Wong [19]	In both FSNet and PFSNet methods, a fuzzy value is assigned to each gene for each sample and then majority voting is used to determine important genes. The codes are available online. The proposed methods only consider the gene grouping membership information.	Bi-level selection

Note: Bilevel selection algorithms are regarded as a special case of pathway-guided gene selection algorithms.
[1]Can be loosely categorized into the indicated category (e.g., stepwise forward).

modules and the selection of relevant features are usually carried out on the same dataset.

Therefore, a through evaluation on which pathways are used during data analysis is highly desirable, in order to maximize the information extraction and to infer true biological meaning.

5. Potential Research Area

So far, the feature selection algorithms we have talked are mainly for cross-sectional data in which data were collected at a single time point. The number of feature selection algorithms for longitudinal data in which the subjects were followed up across time and the corresponding data were collected at different time points is not comparable to that of cross-selection data. To name several, the EDGE method [63, 64], the Generalized Estimating Equation- (GEE-) based screening procedure by [65], the penalized-GEE method [66], and the Penalized-GEE with Grid Search (PGS) method by [67] are included in this small-sized list of longitudinal feature selection algorithms.

As far as the pathway-based feature selection algorithms are considered, to the best of our knowledge, one of our extensions to the SAMGSR method [10], the two-level SAMGSR method, is the only approach that incorporates pathway information to specifically deal with longitudinal data [68]. In the two-level SAMGSR method, the reduction step of the SAMGSR algorithm [10] is applied twice hierarchically. Specifically, the selected gene sets are further reduced to their respective important components, i.e., genes, and then the important time points in selected genes are identified subsequently. Nevertheless, the two-level SAMGSR only considers the grouping membership information. The results of several real-world applications where the diseases under investigation include non-small-cell lung cancer, multiple sclerosis, and traumatic injury [36, 68, 69] have suggested the performance improvement for a pathway-guided method only considering the grouping information over a conventional method may be trivial. In contrast, when a pathway-based method accounts for extra pathway knowledge such as the connectivity information among genes and regulation direction recording which genes regulate which genes, its performance might be promoted dramatically.

One major finding of our previous studies [68, 70] is that the gene expression profiles across different time points may be regarded as a gene set and then some suitable pathway analysis methods may be adopted to select relevant genes for longitudinal data. In the light of this, summary scores at the pathway level such as means, medians [71], the first principal components (PCs) [72], and the sign averages [17, 73] which average out the signed expression values, with signs indicating the association directions between genes and outcome, or more statistically complicated ones like the pathway deregulation scores (PDS) [74], may be chosen

Table 2: Penalty terms used in the penalty methods.

Methods	Mathematical notation	Characteristics
Li & Li, 2008 [27]	$\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)$ Here, d_u is the degree of freedom for gene u, recording the sum of weights for all genes connected to gene u. $w(u,v)$ is the weight for the edge between genes u and v.	Aims at smoothing the β coefficients over the network, ignoring that neighboring genes might have β's in opposite directions.
[55]	$\sum_{u\sim v}\left(\frac{\text{sign}(\widetilde{\beta_u})\beta_u}{\sqrt{d_u}}-\frac{\text{sign}(\widetilde{\beta_v})\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)$ Here, $\widetilde{\beta_u}$ is the estimated value of β coefficient for gene u, and sign (x) represents the sign of x, if x>0 sign(x)=1; x<0 sign(x)=-1; otherwise sign(x)=0.	Accounts for that two connected genes might have β's with different signs, but may not work well since it is difficult to estimate the signs for β's.
[30]	$\sum_{u\sim v}\left[\left(\frac{\beta_u}{\sqrt{d_u}}\right)^\gamma+\left(\frac{\beta_v}{\sqrt{d_v}}\right)^\gamma\right]^{1/\gamma} w(u,v),\ \textit{and}\ \gamma>1$	Shrinks the weighted β's of two neighboring genes towards each other, but the estimates may be severely biased.
[26, 56]	for $\gamma=\infty$, it becomes $\sum_{u\sim v}\max\left(\frac{\lvert\beta_u\rvert}{\sqrt{d_u}},\frac{\lvert\beta_v\rvert}{\sqrt{d_v}}\right)w(u,v)$	A 2-step procedure is used to reduce biases; it is proved that this performs better than that with smaller γ
[57]	$\sum_{u\sim v}\left\lvert I\left(\frac{\lvert\beta_u\rvert}{d_u}\neq 0\right)-I\left(\frac{\lvert\beta_v\rvert}{d_v}\neq 0\right)\right\rvert w(u,v)$ Here, I (x) is an indicator. If the condition x is true I(x)=1, otherwise its value is 0.	Encourages simultaneous selection of neighboring genes in the network. But the Indictor function I is not continuous and thus needs special care.
The generalize elastic net: [29]	$\lambda_1\sum_u D_u\lvert\beta_u\rvert+\frac{\lambda_2}{2}\beta^T P\beta$ Here D and P are additional penalty weights for individual genes (gene-level penalty) and gene pairs (pathway-level penalty).	Includes the network-constrained penalty term by [27] as a special case, capable of accommodating any positive semi-definite measure of dissimilarity between pairs of genes.

Table 3: Penalty terms used in the bilevel selection methods.

Methods	Mathematical notation	Characteristics
Group LASSO [58]	General form $f_o(\sum_{k=1}^{K_j} f_I(\lvert\beta_{j,k}\rvert))$ See equation (2) for what f_o, f_I and $\beta_{j,k}$ represent	It cannot identify the important genes within the selected gene sets and thus is actually incapable of bilevel selection and also heavily shrinks large coefficients (leading to estimate biases for large coefficients)
Group bridge [59]	Outer bridge penalty + inner ridge penalty	It can provide sparse solutions at both pathway and gene levels, but it is associated with big empirical difficulties since the bridge penalty is not everywhere differentiable.
Group MCP [31]	Outer bridge penalty+ inner LASSO penalty	Allow coefficients to grow large and groups to remain sparse.
Group exponential LASSO [18]	Outer MCP penalty+ inner MCP penalty	A decay parameter controls the degree to which gene selection is coupled together within gene sets and has several advantages over the other composite penalty term such as group bridge.
Sparse group LASSO [60]	Outer exponential penalty + inner LASSO penalty	Convex and thus highly likely to get the global minimum, but extra care is needed since the group coordinate descent algorithms cannot be applied.
	$\lambda_1 \sum_j \sum_k \lvert\beta_{j,k}\rvert + \lambda_2 \sum_{j=1}^{J} \lVert\beta_j\rVert_2$ taking the additive format	

Note: the general formatting for group LASSO, group bridge, and group MCP was given by Breheny & Huang [31]. It is too general to guarantee all combinations of outer and inner penalties produce sensible models. Thus the second general form was proposed by Huang et al. [59] to address this issue specifically.

to generate pseudo genes as representatives for respective pathways, and then a longitudinal feature selection process has been downgraded to a classic feature selection process.

Furthermore, one may be also interested in finding those monotonically changed genes as the disease progresses, which may be regarded as a special case of the feature selection for longitudinal data. The word "monotonic" means descending or ascending change patterns across time or stages/grades. To the best of our knowledge, no pathway-based algorithms have been proposed to tackle this specific topic. Therefore, more investigation is warranted to explore if a pathway-guided method is superior to a conventional method such as [75] in selecting monotonic genes. In summary, pathway-guided gene selection may play more roles on identifying potential biomarkers for longitudinal omics data.

6. Conclusions

In this article, we present a review on pathway-based feature selection algorithms. First, based on to what a feature corresponds, pathway-based feature selection methods are classified into three categories, pathway-level selection methods, bilevel selection methods, and pathway-guided gene selection methods. By focusing on the selection of individual genes where pathway information is incorporated as a prior to guide feature selection, pathway-guided gene selection methods were reviewed and discussed in detail. Additionally, given the importance of penalization in the process of feature selection, the commonly used penalty functions in a pathway-guided gene selection method were reviewed. Last, we point out one potential research area in which pathway-guided gene selection deserves more attention, namely, longitudinal gene expression data analysis.

We believe this review provides valuable insights for computational biologists/biostatisticians and stimulates them to develop more elegant pathway-guided gene selection algorithms. The development and wide application of such algorithms to reveal underlying pattern, elucidate the etiology and progression of complex diseases, and guide more "personalized" treatment strategies will contribute substantially to make biology more computable.

Conflicts of Interest

No conflicts of interest have been declared.

Authors' Contributions

Suyan Tian and Chi Wang designed the study. Suyan Tian, Chi Wang, and Bing Wang wrote the paper. Suyan Tian and Chi Wang participated in the critical reviewing of the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgments

This study was supported by a fund (no. 31401123) from the Natural Science Foundation of China.

References

[1] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[2] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in Bioinformatics*, vol. 2015, 13 pages, 2015.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[4] I. Dinu, J. D. Potter, T. Mueller et al., "Improving gene set analysis of microarray data by SAM-GS," *BMC Bioinformatics*, vol. 8, p. 242, 2007.

[5] D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. E. Visvader, and G. K. Smyth, "ROAST: Rotation gene set tests for complex

microarray experiments," *Bioinformatics*, vol. 26, no. 17, pp. 2176–2182, 2010.

[6] P. Creixell, J. Reimand, S. Haider et al., "Pathway and network analysis of cancer genomes," *Nature Methods*, vol. 12, no. 7, pp. 615–621, 2015.

[7] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS Computational Biology*, vol. 8, no. 2, Article ID e1002375, 2012.

[8] M. Bayerlová, K. Jung, F. Kramer, F. Klemm, A. Bleckmann, and T. Beißbarth, "Comparative study on gene set and pathway topology-based enrichment methods," *BMC Bioinformatics*, vol. 16, p. 334, 2015.

[9] Q. Liu, I. Dinu, A. J. Adewale, J. D. Potter, and Y. Yasui, "Comparative evaluation of gene-set analysis methods," *BMC Bioinformatics*, vol. 8, p. 431, 2007.

[10] I. Dinu, J. D. Potter, T. Mueller et al., "Gene-set analysis and reduction," *Briefings in Bioinformatics*, vol. 10, no. 1, pp. 24–34, 2009.

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

[12] J.-H. Hung, T.-H. Yang, Z. Hu, Z. Weng, and C. DeLisi, "Gene set enrichment analysis: Performance evaluation and usage guidelines," *Briefings in Bioinformatics*, vol. 13, Article ID bbr049, pp. 281–291, 2012.

[13] M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus, "Pathway analysis: state of the art," *Frontiers in Physiology*, vol. 6, pp. 1–16, 2015.

[14] A. L. Tarca, G. Bhatti, and R. Romero, "A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity," *PLoS ONE*, vol. 8, no. 11, 2013.

[15] M. K. Jaakkola and L. L. Elo, "Empirical comparison of structure-based pathway methods," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 336–345, 2016.

[16] S. Tian, C. Wang, H. H. Chang, and J. Sun, "Identification of prognostic genes and gene sets for early-stage non- small cell lung cancer using bi-level selection methods," *Scientific Reports*, pp. 1–8, 2017.

[17] K. H. Eng, S. Wang, W. H. Bradley, J. S. Rader, and C. Kendziorski, "Pathway index models for construction of patient-specific risk profiles," *Statistics in Medicine*, vol. 32, no. 9, pp. 1524–1535, 2013.

[18] P. Breheny, N. Riverside, N. Cphb, and I. City, "The group exponential lasso for bi-level variable selection," *Biometrics*, vol. 71, pp. 731–740, 2015.

[19] K. Lim and L. Wong, "Finding consistent disease subnetworks using PFSNet," *Bioinformatics*, vol. 30, no. 2, pp. 189–196, 2014.

[20] S. Tian, H. H. Chang, and C. Wang, "Weighted-SAMGSR: Combining significance analysis of microarray-gene set reduction algorithm with pathway topology-based weights to select relevant genes," *Biology Direct*, vol. 11, p. 50, 2016.

[21] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[22] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, pp. 1–10, 2007.

[23] M. Johannes, J. C. Brase, H. Fröhlich et al., "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients," *Bioinformatics*, vol. 26, no. 17, pp. 2136–2144, 2010.

[24] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, "GeneRank: Using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, p. 233, 2005.

[25] A. Zhang and S. Tian, "Classification of early-stage non–small cell lung cancer by weighing gene expression profiles with connectivity information," *Biometrical Journal*, vol. 60, no. 3, pp. 537–546, 2018.

[26] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," *BMC Bioinformatics*, vol. 10, Suppl I, p. S21, 2009.

[27] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[28] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines," *BMC Systems Biology*, vol. 5, p. 161, 2011.

[29] A. Sokolov, D. E. Carlin, E. O. Paull, R. Baertsch, and J. M. Stuart, "Pathway-based genomics prediction using generalized elastic net," *PLoS Computational Biology*, vol. 12, no. 3, pp. 1–23, 2016.

[30] W. Pan, B. Xie, and X. Shen, "Incorporating predictor network in penalized regression with application to microarray data," *Biometrics*, vol. 66, no. 2, pp. 474–484, 2010.

[31] P. Breheny and J. Huang, "Penalized methods for bi-level variable selection," *Statistics and Its Interface*, vol. 2, no. 3, pp. 369–380, 2010.

[32] N. Bandyopadhyay, T. Kahveci, S. Goodison, Y. Sun, and S. Ranka, "Pathway-based feature selection algorithm for cancer microarray data," *Advances in Bioinformatics*, vol. 2009, Article ID 532989, 16 pages, 2009.

[33] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, Article ID e1000217, 2008.

[34] A. Razi, F. Afghah, S. Singh, and V. Varadan, "Network-based enriched gene subnetwork identification?: a game-theoretic approach," *Biomedical Engineering and Computational Biology*, vol. 7, pp. 1–14, 2016.

[35] Q. Zhang, J. Li, D. Wang, and Y. Wang, "Finding disagreement pathway signatures and constructing an ensemble model for cancer classification," *Scientific Reports*, pp. 1–11, 2017.

[36] L. Zhang, L. Wang, B. Du, T. Wang, P. Tian, and S. Tian, "Classification of non-small cell lung cancer using significance analysis of microarray-gene set reduction algorithm," *BioMed Research International*, vol. 2016, Article ID 2491671, 8 pages, 2016.

[37] N. Doungpan, W. Engchuan, J. H. Chan, and A. Meechai, "GSNFS: Gene subnetwork biomarker identification of lung cancer expression data," *BMC Medical Genomics*, vol. 9, 2016.

[38] W. Engchuan, A. Meechai, S. Tongsima, N. Doungpan, and J. H. Chan, "Gene-set activity toolbox (GAT): a platform for microarray-based cancer diagnosis using an integrative gene-set analysis approach," *Journal of Bioinformatics and Computational Biology*, Article ID 1650015, 2016.

[39] W. H. Chan, M. S. Mohamad, S. Deris et al., "Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme," *Computers in Biology and Medicine*, vol. 77, pp. 102–115, 2016.

[40] C. Li, X. Li, and Y. Miao, "SubpathwayMiner: a software package for flexible identification of pathways," *Nucleic Acids Research*, vol. 37, 2009.

[41] H. Parkinson, U. Sarkans, N. Kolesnikov et al., "Arrayexpress update-An archive of microarray and high-throughput sequencing-based functional genomics experiments," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1002–D1004, 2011.

[42] Y. Cun and H. F. Fröhlich, "Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions," *BMC Bioinformatics*, vol. 13, 2012.

[43] C. Staiger, S. Cadot, R. Kooter et al., "A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer," *PLoS ONE*, vol. 7, no. 4, 2012.

[44] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[45] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[46] D. Croft, A. F. Mundo, and R. Haw, "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, no. 1, pp. D472–D477, 2014.

[47] N. Auslander, A. Wagner, M. Oberhardt, and E. Ruppin, "Data-driven metabolic pathway compositions enhance cancer survival prediction," *PLoS Computational Biology*, pp. 1–17, 2016.

[48] D. Henriques, A. F. Villaverde, M. Rocha, J. Saez-Rodriguez, and J. R. Banga, "Data-driven reverse engineering of signaling pathways using ensembles of dynamic models," *PLoS Computational Biology*, vol. 13, no. 2, Article ID e1005379, 2017.

[49] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, 2008.

[50] A. A. Margolin, I. Nemenman, K. Basso et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, p. S7, 2006.

[51] S. Yang, L. Yuan, Y. Lai, X. Shen, P. Wonka, and J. Ye, "Feature grouping and selection over an undirected graph," *KDD*, pp. 922–930, 2012.

[52] Y. Zhu, X. Shen, and W. Pan, "Simultaneous grouping pursuit and feature selection over an undirected graph," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 713–725, 2013.

[53] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang, "Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment," *PLoS Computational Biology*, vol. 9, no. 3, Article ID e1002975, 2013.

[54] X. Wu, H. Tang, A. Guan, F. Sun, H. Wang, and J. Shu, "Finding gastric cancer related genes and clinical biomarkers for detection based on gene-gene interaction network," *Mathematical Biosciences*, vol. 276, pp. 1–7, 2016.

[55] C. Li and H. Li, "Variable selection and regression analysis for graph-structured covariates with an application to genomics," *The Annals of Applied Statistics*, vol. 4, no. 3, pp. 1498–1516, 2010.

[56] C. Luo, W. Pan, and X. Shen, "A Two-Step Penalized Regression Method with Networked Predictors," *Statistics in Biosciences*, vol. 4, no. 1, pp. 27–46, 2012.

[57] S. Kim, W. Pan, and X. Shen, "Network-based penalized regression with application to genomic data," *Biometrics: Journal of the International Biometric Society*, vol. 69, no. 3, pp. 582–593, 2013.

[58] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[59] J. Huang, P. Breheny, and S. Ma, "A Selective review of group selection in high-dimensional models," *Statistical Science*, vol. 27, no. 4, pp. 481–499, 2012.

[60] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.

[61] S. Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in Bioinformatics*, pp. 1–18, 2017.

[62] E. A. Serin, H. Nijveen, H. W. Hilhorst, and W. Ligterink, "Learning from co-expression networks: Possibilities and challenges," *Frontiers in Plant Science*, vol. 7, 2016.

[63] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: Extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, no. 4, pp. 507-508, 2006.

[64] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12837–12842, 2005.

[65] P. Xu, L. Zhu, and Y. Li, "Ultrahigh dimensional time course feature selection," *Biometrics: Journal of the International Biometric Society*, vol. 70, no. 2, pp. 356–365, 2014.

[66] L. Wang, J. Zhou, and A. Qu, "Penalized generalized estimating equations for high-dimensional longitudinal data analysis," *Biometrics*, vol. 68, no. 2, pp. 353–360, 2012.

[67] Y. Zheng, Z. Fei, W. Zhang et al., "PGS: A tool for association study of high-dimensional microRNA expression data with repeated measures," *Bioinformatics*, vol. 30, no. 19, pp. 2802–2807, 2014.

[68] S. Tian, C. Wang, and H. H. Chang, "To select relevant features for longitudinal gene expression data by extending a pathway analysis method," *F1000Research*, vol. 7, p. 1166, 2018.

[69] L. Zhang, L. Wang, P. Tian, S. Tian, and K. Brusgaard, "Identification of genes discriminating multiple sclerosis patients from controls by adapting a pathway analysis method," *PLoS ONE*, vol. 11, no. 11, Article ID 0165543, p. e0165543, 2016.

[70] S. Tian, C. Wang, and H. H. Chang, "A longitudinal feature selection method identifies relevant genes to distinguish complicated injury and uncomplicated injury over time," *BMC Medical Informatics and Decision Making*, vol. 18, no. S5, 2018.

[71] Z. Guo, T. Zhang, X. Li et al., "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics*, vol. 6, p. 58, 2005.

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

[73] S. D. Zhao, G. Parmigiani, C. Huttenhower, and L. Waldron, "Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis," *Bioinformatics*, vol. 30, no. 21, pp. 3062–3069, 2014.

[74] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 110, no. 16, pp. 6388–6393, 2013.

[75] H. Wang, H. Sun, T. Chang et al., "Discovering monotonic stemness marker genes from time-series stem cell microarray data," *BMC Genomics*, vol. 16, no. Suppl 2, p. S2, 2015.