Expert Systems With Applications 135 (2019) 102–112

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An integrative analysis system of gene expression using self-paced learning and SCAD-Net

Hai-Hui Huang[a,b,*], Yong Liang[b]

[a] School of Information Science and Engineering & Provincial Demonstration Software Institute, Shaoguan University, Shaoguan, China
[b] Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau, China

ARTICLE INFO

Article history:
Received 25 October 2018
Revised 7 May 2019
Accepted 6 June 2019
Available online 7 June 2019

Keywords:
Integrative analysis system
Meta-analysis
Regularization
Variable selection
Gene expression

ABSTRACT

Background: Few proposed gene biomarkers have been satisfactory in clinical applications. That is mainly due to the small studies sample size. Because of the batch effect, different gene-expression studies cannot be merged directly. Many integrative methods have attempted to integrate various datasets to eliminate the batch effect while keeping biological information intact. However, due to the complexity of the batch effect, it cannot be eliminated, and these methods may even add new systematic errors to the data, further complicating integrated data. Therefore, direct analysis of the merged data may cause some issues. In this paper, we suggest a novel integrative analysis framework for merged gene-expression data. The framework adopts the self-paced learning. This method allows samples to be automatically added into the training period, from simple to intricate, in a purely self-paced way. Moreover, the framework includes a new feature selection method, the SCAD-Net regularization method, a combination of SCAD and network-based penalties to integrates the biological network knowledge. The simulation shows that the proposed method outperforms the benchmark with more accurate marker identification. The analysis of seven large NSCLC gene expression datasets shows that the proposed method not only results in higher accuracies, but also identifies potential therapeutic markers and pathways in NSCLC. In conclusion, we provide a new and efficient integrative analysis system of gene expression, for the search for new reliable diagnosis or targeted therapy biomarker.

1. Introduction

To date, numerous gene biomarker studies have been completed (Dang et al., 2018; Reis-Filho & Pusztai, 2011). Unfortunately, few of the proposed gene biomarkers are satisfied in clinical applications. That is mainly due to small study sample sizes (Ali et al., 2014; Hay, Thomas, Craighead, Economides, & Rosenthal, 2014). Small sample sizes reduce statistical efficacy, which can result in false conclusions. Sufficient sample is required to produce effective statistical analysis and valid conclusions.

The increasing amount and availability of large gene expression studies motivate the development of integrative analysis that combines multiple datasets or relevant results. However, although some gene expression studies share the same goal, constituent datasets have typically been generated using diverse processing facilities, different data platforms and return expression values on different numerical scales (often called the batch effect). Therefore, merging information from different gene expression studies poses a statistical challenge.

Extensive efforts have been made to address this challenge and can be divided into two distinct approaches: meta-analysis and integrative analysis via data merging (Ma, 2009). The first approach, meta-analysis, uses statistical methods that combining results from different studies. However, meta-analysis is trivial and several conditions are critical for viable results, and small violations of those conditions can lead to misleading results (Walker, Hernandez, & Kattan, 2008). The second approach is the integrative analysis method, which merges diverse datasets into a union dataset, and performs analysis based on this newly integrated dataset. Its main advantage over meta-analysis is higher result statistical significance due to large datasets (Lazar et al., 2013). Many methods have been proposed based on this idea, such as, Distance-weighted discrimination (DWD) (Benito et al., 2004), a method that seeks to identify separating hyperplane that maximizes the separation

Abbreviations: NSCLC, non-small cell lung cancer; SCAD-Net, SCAD penalized network-based regularization, SCAD-NL: SCAD-Network-based penalized logistic regression model; SPS-Net, SPL-SCAD-Network-based regularization; SPS-NL, SPL-SCAD-Network-based penalized logistic regression model.

* Corresponding author at: School of Information Science and Engineering & Provincial Demonstration Software Institute, Shaoguan University, Shaoguan, China.
E-mail addresses: tomyhwang@163.com (H.-H. Huang), yliang@must.edu.mo (Y. Liang).

https://doi.org/10.1016/j.eswa.2019.06.016
0957-4174/

between each sample from the different classes, and then moves systematic bias along the normal direction vector until its mean distance attains the hyperplane. Empirical Bayes (EB, also named COMBAT) (Johnson, Li, & Rabinovic, 2007), is an approach that establishes a mixed-effect correction for each gene, and estimates the correction by merging information from multiple genes with similar expression traits in each batch. Cross-platform normalisation (XPN) (Shabalin, Tjelmeland, Fan, Perou, & Nobel, 2008), is a procedure that identifies blocks of gene and samples in multiple datasets with homogeneous expression traits. PLIDA (Deshwar & Morris, 2014), this method uses topic models to summarize the expression patterns in each dataset before normalizing the topics learned from each dataset using per-gene multiplicative weights. WaveICA (Deng et al., 2019), this strategy uses the time trend of samples over the injection order, decomposes the original data into multi-scale data with different features, extracts and removes the batch effect information in multi-scale data, and gets clean data. These integrative methods seek to combine various datasets into an integrated dataset to eliminate the batch effect while keeping biological information intact. However, given the complicated sources of the batch effect, it cannot be eliminated, and these methods may even add new systematic errors to the data, further complicating the integrated data. Thus, directly analyzing the data may cause some issues (Lazar et al., 2013; Qi et al., 2016). Therefore, a new learning strategy is needed to adapt to this situation.

Recently, a novel learning strategy called self-paced learning (SPL), was proposed (Bengio, Louradour, Collobert, & Weston, 2009; Kumar, Packer, & Koller, 2010). This strategy begins with simple concepts and builds up to more complex ideas. SPL can adaptively recognize easy and hard samples based on what the model has already learned, with increasingly more difficult samples used for model training. The SPL strategy has been successfully used in various machine learning problems (Jiang, Meng, Mitamura, & Hauptmann, 2014). Furthermore, some convergence properties of SPL have also been a discussion (Z. Ma et al., 2018), makes the SPL more theoretically rational. Therefore, SPL is a promising learning mechanism that will help us build a more accurate prediction model for integrated gene-expression datasets.

An efficient feature selection method is needed to better identify disease-related biomarkers from tens of thousands of gene features. The regularization method is commonly used for feature selection. It enforces small coefficients to 0 and therefore presents a sparse representation of the result. Many regularization procedures have been recommended for gene selection, including the Lasso (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) technique (Fan & Li, 2001), the Elastic net (Zou & Hastie, 2005) method, the adaptive Lasso approach (Zou, 2006) and Hybrid L1/2 + 2 Regularization (Huang & Liang, 2018; Huang, Liu, & Liang, 2016; Liang et al., 2013). However, these methods lack a built-in mechanism to fuse prior biological information regarding genes that are frequently available in scientific applications. Integrating biological network information with an analysis of gene expression data has provided useful prior knowledge for the removal of noise and detection of confounding factors from genomics data for many regression and classification models (Huang, Liang, & Liu, 2015; Li & Li, 2008).

The complexity of biological data and the suggestions above have prompted us to propose a new integrative analysis system or framework (Fig. 1).

In this framework, different gene-expression datasets are integrated into a single unified dataset using a popular integration approach. Then, the SPL-SCAD-Network-based regularization (SPS-Net) method is coupled with a logistic regression model to fit the data for biomarker selection. More specifically, the SPS-Net consists of three parts: 1) *The SCAD penalty*. The *SCAD* penalty is applied to enforce model sparsity. This penalty offers unbiased estimates for large coefficients. Also, model estimates by the SCAD method have valuable theoretical properties, for example, Oracle (if a right sub-model were known) (Fan & Li, 2001); 2) *The network-based penalty*. We apply a network-based penalty (or quadratic Laplacian penalty) to enforce smoothness between the coefficients of neighboring genes on a given gene regulatory network; 3) *The self-paced learning (SPL) method*. We integrate the SPL regime into the model training, and this technique prompts the use of easy samples (high confidence samples) first and increasingly guide the learning algorithm to more complex samples (low confidence samples). This idea is crucial to integrative gene-expression data analysis, as this data often has heavy noises and outliers.

We applied the proposed framework to seven public NSCLC datasets for performance testing. The outcomes of the experiment indicate that the framework could be useful in identifying a set of robust disease-related gene signatures.

The rest of the paper is organized as follows. Part 2 presents a penalized logistic regression model. Then, the SCAD-network-based regularization method is proposed, and some theorems of this new penalty are also discussed. Then, we present the SPL and combine this learning strategy with the SCAD-network-based penalty, and couple with a logistic regression model to create the final model. In Part 3, coefficient estimators of the SCAD-network-based penalty are derived and we propose an efficient algorithm for solving the final model. In Part 4, we evaluate the performance of our proposed method through a comprehensive simulation analysis and real mRNA expression level data experiment. A brief discussion and conclusion are presented in Part 5.

2. Method

2.1. SCAD Network-based penalty

Suppose that dataset D has n samples $D=\{(X_1,y_1),(X_2,y_2),\ldots,(X_n,y_n)\}$, where $X_i=(x_{i1},x_{i2},\ldots,x_{ip})$ is the ith sample with p genes and y_i is the corresponding dependent variable that consist of a binary value of either 0 or 1. Define a classifier $f(x)=e^x/(1+e^x)$ and the logistic regression is defined as:

$$P(y_i=1|X_i) = \left(X_i'\beta\right) = \frac{exp\left(X'_i\beta\right)}{1+exp(X'_i\beta)} \tag{1}$$

where $\beta=(\beta_1,\ldots,\beta_p)$ are the estimated coefficients. Using simple algebra, the regression model can be presented as:

$$l(\beta) = -\sum_{i=1}^{n}\left\{ y_i log\left[f\left(X_i'\beta\right)\right] + (1-y_i)\log\left[1-\left(X_i'\beta\right)\right]\right\} \tag{2}$$

However, in most gene expression studies, the number of genes typically far exceeds the sample size. This situation refers to as a high-dimensional and low sample size problem, and standard logistic regression method cannot be directly used to estimate the regression parameters. The regularization method is one of the popular techniques to resolve the issue of high dimensionality, and can be expressed as:

$$L(\lambda,\beta) = l(\beta) + P(\beta),$$

where $P(\beta)$ represents the regularization term. A popular regularization term is the L_1 (Lasso) method (Tibshirani, 1996) which has the penalty function $P_{\lambda,Lasso}(\beta)=\lambda\sum_{j=1}^{p}|\beta_j|^1$, where λ is any non-negative value. As a result of the singularity of the L_1 penalty function, a L_1 penalized logistic model automatically selects features by shrinking small coefficients to zero. However, when λ is too big, the estimation of large β may suffer from substantial bias, and if λ is too small, the solution may not be sufficiently sparse. To

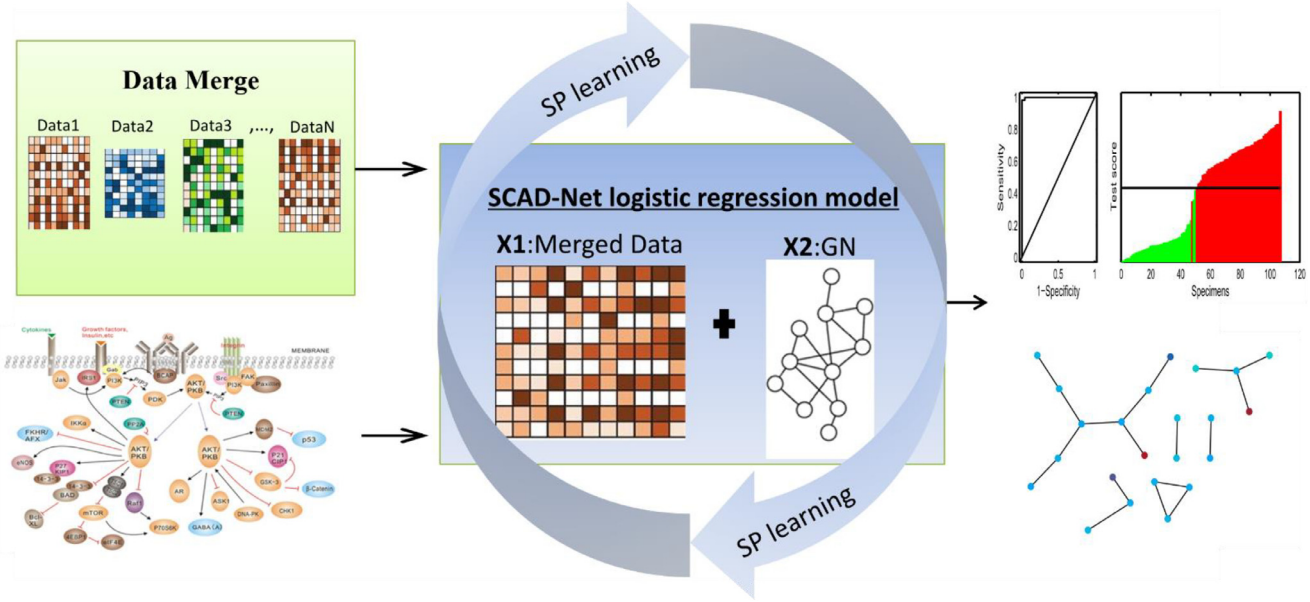

Fig. 1. Overview of the proposed integrative analysis framework.

overcome this issue, Fan and Li (Fan & Li, 2001) proposed the SCAD penalty, defined as:

$$P_{\lambda,SCAD}(\beta) = \begin{cases} \lambda|\beta|, & if\ 0 \le |\beta| < \lambda, \\ -\frac{\beta^2 - 2\alpha\lambda|\beta| + \lambda^2}{2(\alpha-1)}, & if\ \lambda \le |\beta| < \alpha\lambda, \\ \frac{(\alpha+1)\lambda^2}{2}, & otherwise, \end{cases} \tag{3}$$

where α is a constant value larger than 2. To deal with the high-correlation situation, Zou and Hastie (Zou & Hastie, 2005) proposed the Elastic net method $P_{\lambda_1,\lambda_2,enet}(\beta) = \sum_{j=1}^{p}(\lambda_1|\beta_j|^1 + \lambda_2|\beta_j|^2)$. Zeng and Xie (Zeng & Xie, 2012) proposed the SCAD-L_2 method which combines the SCAD and L_2 penalties. Such methods achieves the grouping effect, where strongly correlated genes tend to be in or out of the result together. However, these approaches were proposed using purely computational points without any prior biological information.

Prior information on gene regulatory interactions is valuable for decoding modular gene patterns. The network-based penalty has been proposed to utilize these prior network knowledge for many applications. For example, Li and Li (Li & Li, 2008), Chen et al. (Chen, Zhang, & C., 2016) and Wang et al. (Wang et al., 2018) recommend a L_1 penalized network-constrained regularization procedure for feature selection and regression analysis of genomic data. In these researches, the network-based function is defined similarly as a quadratic form of the Laplacian matrix connected with the genes interaction network. As we argue above, the L_1 penalty suffers additional bias and may not be sufficiently sparse in some situations. Compared with the L_1 penalty, the SCAD approach avoids excessive penalties on large coefficients and induces the oracle property. Therefore, it is reasonable to adopt the SCAD method instead of the L_1 penalty. Here, we propose a SCAD penalized network-based (SCAD-Net) method. It can be formulated as follows:

$$P_{\lambda_1,\lambda_2,SCAD-Net}(\beta) = P_{\lambda_1,SCAD}(\beta) + \lambda_2\beta L\beta, \tag{4}$$

where L represents the symmetric Laplacian matrix, which integrates the biological network knowledge, and the $\beta L\beta$ enforce a smooth result of β on the network. Eq. (4) can be rewritten as:

$$P_{\lambda_1,\lambda_2,SCAD-Net}(\beta) = P_{\lambda_1,SCAD}(\beta) + \lambda_2 \sum_{1\le i<k\le p;} w_{ik}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}}\right)^2, \tag{5}$$

where gene i and gene k are linked, then $w_{ik} = 1$ or a value ranging from 0 to 1, else $w_{ik} = 0$; d_i and d_k represent the degrees of genes i and k respectively, meaning the number of edges linked with i (or k); λ_1 and λ_2 adjust the sparsity and smoothness of the model respectively. Then, the SCAD penalized network-based logistic regression model (SCAD-NL) is defined as:

$$\hat{\beta} = argmin_{\beta}\left\{ l(\beta) + P_{\lambda_1,SCAD}(\beta) + \lambda_2 \sum_{1\le i<k\le p;} w_{ik}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}}\right)^2 \right\}, \tag{6}$$

where the first term is the logistic regression loss function, resulting in a classification prediction model. The second term is the SCAD penalty, which ensures sparsity, allowing the solutions to have better biological interpretations. The last term is the network-based penalty, which captures critical prior knowledge, and makes the connected genes in the network to be smoothed-regression coefficients.

2.2. The grouping effect of the SCAD-NL

In this section, we show the SCAD-NL share grouping effect. The proofs of following Lemma and Theorem are provided in the Supplement File except for a short statement for Theorem 1.

Lemma 1 Assume $P_{j,\lambda_1,\lambda_2}(\beta)$ is the SCAD-Net function of the single feature β_j, with the remaining items of β fixed. For the SCAD-Net penalty with $\lambda_2 > \frac{1}{(2(a-1))}$, $P_{j,\lambda_1,\lambda_2}(\beta)$ is a convex function of β_j for all j.

The following Theorem, which is the explicit argument from the Lemma 2 of (Zou & Hastie, 2005) since the SCAD-NL is a convex function, ensure the grouping effect for a situation when two predictor variables are equal.

Theorem 1. Suppose that $\hat{\boldsymbol{\beta}}$ is calculated by Eq. (6), and suppose that $\boldsymbol{x_j} = \boldsymbol{x_i}$, then we have $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ for any $\boldsymbol{\lambda_2} > \frac{1}{(2(\boldsymbol{a}-1))}$.

The following Theorem demonstrates a quantitative description of the grouping effect of the SCAD-NL.

Theorem 2. *Suppose that* $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$. $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} > 0$ *and* $\boldsymbol{\lambda}_2 > \frac{1}{(2(\boldsymbol{a}-1))}$. *Define*

$$D(i,j) = \frac{\left|\hat{\beta}_i - \hat{\beta}_j\right|}{|y|_1}$$

Then,

$$D(i,j) \le \frac{1}{\left(2\lambda_2 - \frac{1}{a-1}\right)}\sqrt{2(1-\rho)}.$$

where $\rho = x_i^T x_j$ is the sample correlation.

Theorem 2 shows the upper bound of the difference between the coefficients of two genes. If ρ is approach to 1, then the Theorem 2 ensures that the coefficients of the two genes are very close.

2.3. Self-paced learning

Although the regularization method plays a crucial role in the gene-expression data analysis especially gene selection, only a few of the selected biomarkers are used in clinical applications. This issue is mainly because these studies are based on small sample size data, which reduces the validity of the conclusions. Several proposals have been proposed to solve this issue by integrating various datasets into an integrated dataset to generate sufficient sample sizes (Benito et al., 2004; Johnson et al., 2007; Shabalin et al., 2008). However, these integration methods cannot eliminate internal bias, and may even add a new random noise and estimation errors to the dataset, reducing the statistical power of the integrative analysis (Qin, Huang, & Begg, 2016). A learning strategy involves learning from the low-level noise samples first to develop a basic or universal biological knowledge, and when it is "powerful" enough, then learning the high-level noise samples to improve knowledge structure. This strategy may substantially increase the statistical power of integrative analysis.

Self-paced Learning (SPL) or curriculum learning (CL) was first suggested in (Bengio et al., 2009; Kumar et al., 2010), inspired by human learning mechanism. Kumar et al. (Kumar et al., 2010) showed the SPL could be a concise optimization model by introducing a penalty term. A typical SPL framework can be expressed as:

$$E(\beta, V) = \sum_{i=1}^{n}\{v_i l(y_i, d(x_i,\beta)) + f(v_i;\tau)\}, \tag{7}$$

where $v_i l(y_i,d(x_i,\beta))$ is a weighted loss term for every sample, $l(y_i,d(x_i,\beta))$ represents a specific loss function; $d(x_i,\beta)$ is the decision function, $V=(v_1,v_2,\ldots v_n)$ is a weight vector for the whole sample set; τ represents an age variable for adjusting the learning pace, and $f(v_i;\ \tau)$ is the SP penalty influence on the sample weight. A common SP penalty is the original hard regularization function $f(v;\ \tau) = -\tau v$, and some variants of this function are discussed in (Meng, Zhao, & Lu, 2017).

By solving the weight vector V with increasing age parameter, SPL allows more samples to be automatically added into the training period, from simple to intricate, in a purely self-paced way.

2.4. Combine the SPL mechanism with the SCAD-NL model

To increase robustness and accuracy in integrative analysis, we combined the SPL mechanism with the SCAD-NL model (SPS-NL), that is

$$\min_{\beta,\mathrm{v}\in[0,1]^n} S(\beta,\mathrm{v};\ \tau,\ \lambda_1,\lambda_2) = \sum_{i=1}^{n}\{v_i l(\beta) + f(v_i;\tau)\} + P_{\lambda_1,\lambda_2,SCAD-Net}(\beta) \tag{8}$$

where the first term is the weighted logistic model. The second term $f(v;\ \tau) = -\tau v_i$ is the SP penalty influence on the weight term v_i and the age parameter τ. The age parameter adjusts the learning pace, with smaller values discouraging complex samples into the training process. The last term $P_{\lambda_1,\lambda_2,SCAD-Net}(\beta)$ represents the SCAD-Net regularization term on β.

3. Calculation

3.1. Coefficient estimators of the SCAD-Net

In this section, we present a novel coordinate-wise update form for the SCAD-Net penalty. Consider the regression model:

$$y = \mathrm{x}_1\beta_1 + \mathrm{x}_2\beta_2 + \ldots + \mathrm{x}_p\beta_p,$$

where the response y is predicted by p predictors $x_1, x_2,\ldots,\ x_p$. Without loss of generality, the predictors and response are all normalized and centered. A normal penalized least-squares function can be expressed as:

$$\begin{aligned}\mathcal{L}(\beta) &= \tfrac{1}{2}\|y - X\beta\|^2 + P(\beta)\\ &= \tfrac{1}{2}\left\|y - \hat{y}\right\|^2 + \tfrac{1}{2}\|X\beta - X\beta\|^2 + P(\beta)\\ &= \tfrac{1}{2}\left\|\hat{\beta}_{OLS} - \beta\right\|^2 + \sum_{j=1}^{p} P\left(\beta_j\right)\end{aligned} \tag{9}$$

where $\hat{y} = \mathrm{X}\hat{\beta}_{OLS}$ and $\hat{\beta}_{OLS} = X^T y$ is the ordinary least-squares (OLS) solution.

First partial derivative concerning β_j of Eq. (9) is given by:

$$\frac{\partial}{\partial\beta_j}\mathcal{L}(\beta) = \beta_j - \hat{\beta}_{j,OLS} + p'\left(\beta_j\right),$$

where $\hat{\beta}_{j,OLS}$ is the jth item of $\hat{\beta}_{OLS}$. By setting all first partial derivatives equal to 0, we get the solution with its jth item given by:

$$\hat{\beta}_j = \hat{\beta}_{j,OLS} - p'\left(\beta_j\right).$$

For the L_1 (Lasso) method, its estimator is given by

$$\hat{\beta}_{j,Lasso} = sign\left(\hat{\beta}_{j,OLS}\right)\left(\left|\hat{\beta}_{j,OLS}\right| - \lambda\right)_+,$$

where $(\partial)_+ = \partial$ *if* $\partial \ge 0$ *and* $(\partial)_+ = 0$ otherwise.

The penalty function of SCAD is defined as $P_{j,SCAD}(\beta) = \sum_{j=1}^{p} g_\lambda(\beta_j)$, where $g_\lambda(.)$ is provided in Eq. (2). The jth item of the SCAD estimator is given by:

$$\hat{\beta}_{j,SCAD} = \begin{cases} sign(\hat{\beta}_{j,OLS})\left(\left|\hat{\beta}_{j,OLS}\right| - \lambda\right)_+, & if\left|\hat{\beta}_{j,OLS}\right| \le 2\lambda,\\ \frac{(\alpha-1)\hat{\beta}_{j,OLS} - \alpha\lambda.sign(\hat{\beta}_{j,OLS})}{(\alpha-2)}, & if 2\lambda \le \left|\hat{\beta}_{j,OLS}\right| < \alpha\lambda,\\ \hat{\beta}_{j,OLS}, & otherwise,\end{cases}$$

where $\hat{\beta}_{j,OLS}$ is unbiased when the estimator has an absolute value bigger than $a\lambda$.

Recall $P_{\lambda_1,\lambda_2,SCAD-Net}(\beta) = P_{\lambda_1,SCAD}(\beta) + \lambda_2 \sum_{1\le i<k\le p;} w_{ik}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}}\right)^2$, and it can be rewritten as:

$$\begin{aligned}P_{\lambda_1,\lambda_2,SCAD-Net}(\beta) = P_{\lambda_1,SCAD}(\beta) &+ \lambda_2\sum_{i=1}^{p} w_{ij}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_j}{\sqrt{d_j}}\right)^2\\ &+\lambda_2 \sum_{1\le i<k\le p;i,k\ne j} w_{ik}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}}\right)^2.\end{aligned} \tag{10}$$

Its first derivative concerning β_j is:

$$P\prime_{\lambda_1,\lambda_2,SCAD-Net}(\beta_j)=\begin{cases} sign(\beta_j)-\beta_j+2\lambda_2\beta_j-t, & if 0<\left|\beta_j\right|\le\lambda_1,\\ \frac{\alpha\lambda_1.sign(\beta_j)-\beta_j}{(\alpha-2)}+2\lambda_2\beta_j-t, & if\lambda_1\le\left|\beta_j\right|<\alpha\lambda_1,\\ 2\lambda_2\beta_j-t, & otherwise, \end{cases} \tag{11}$$

where $t=\lambda_2\sum_{i=1}^{p}\frac{w_{ij}\beta_i}{\sqrt{d_i d_j}}$.

By setting the first partial derivatives of the SCAD-Net penalty equal to 0, we get a solution for the naive SCAD-Net estimator with the jth item given by:

$$\hat{\beta}_{j,naive}=\begin{cases} \frac{sign(\hat{\beta}_{j.OLS})(\left|\hat{\beta}_{j.OLS}+t\right|-\lambda_1)_+}{1+2\lambda_2}, & if\left|\hat{\beta}_{j.OLS}\right|\le 2\lambda_1(1+\lambda_2)-t,\\ \frac{(\alpha-1)\hat{\beta}_{j.OLS}-\alpha\lambda_1.sign(\hat{\beta}_{j.OLS})+(\alpha-1)t}{(\alpha-2)(1+2\lambda_2)-1}, & if 2\lambda_1(1+\lambda_2)-t\le\left|\hat{\beta}_{j.OLS}\right|\\ \frac{\hat{\beta}_{j.OLS}+t}{1+2\lambda_2}, & <\alpha\lambda_1(1+2\lambda_2)-t,\\ & otherwise. \end{cases} \tag{12}$$

When $|\hat{\beta}_{j.OLS}|$ has a large value, $\hat{\beta}_{j.\ naive}$ is a biased estimator. Thus, to obtain an unbiased solution, the final SCAD-Net estimator is defined as rescale:

$$\hat{\beta}_j=(1+2\lambda_2)\hat{\beta}_{j.\ naive}-t. \tag{13}$$

3.2. Solution of the SPS-NL model

We use the alternate optimization search algorithm to optimize the SPS-NL model. Supposing we have training dataset$X_{n\times p}$ and its corresponding binary dependent variables y_n, the detailed optimization procedure is shown as follows:

Initialize. Some optimization variables and parameters are initialized in this process.

Remark: Note that the samples used in the first round $V^o=(v_1,\ldots,v_n)$ are crucial to the success of the framework. The following procedures are adopted to select the confidential samples used in the startup.

1) 70% samples of the training set$X$are randomly selected and denoted as E_i.
2) The classifier is trained based on E_i using the SCAD-NL method, and applied to the training set$X$using a proper cut-off point to determine a predicted classification result $\hat{y}^{(i)}$.
3) The above procedures are repeated n times, to generate $\hat{y}^{(1)},\hat{y}^{(2)},\ldots,\hat{y}^{(n)}$.
4) Let $P=y.*,\hat{y}^{(1)}.*\hat{y}^{(2)}.*,\ldots,*\hat{y}^{(n)};F=\sim y.*\sim\hat{y}^{(1)}.*\sim\hat{y}^{(2)}.*,\ldots,.*\sim\hat{y}^{(n)}$, where .* is dot product operator and $\sim$ is the NOT logical operator.
5) $K=P+F$;

Set $V^o=K$, and τ is set to a small value to promoted easy samples in the first round of training.

Fix V and update β^t. When V is fixed, the SPS-NL problem is solved as a standard SCAD-NL problem:

$$\beta^t=argmin_\beta\left\{l(\beta^{t-1})+P_{\lambda_1,\lambda_2,SCAD-Net}(\beta^{t-1})\right\}.$$

With the coefficient estimators of SCAD-Net as devised in Eq. (13) and the coordinate descent algorithm (CDA) (Friedman, Hastie, & Tibshirani, 2010; H.-H. Huang et al., 2016; H. H. Huang, Liu, Li, & Liang, 2017), we can easily solve the SCAD-NL model.

Fix β and update v_i^t. The mechanism of this process measures the sample "quality" according to v_i. By calculating the first derivative with respect to v_i of SPS-NL, we have:

$$\frac{\partial S}{\partial v_i}=l\left(y_i,x_i^T\beta^t\right)-\tau$$

With simple algebra, the closed-form updating equation for v_i is given by:

$$v_i^t=\begin{cases}1, & l_i\le\tau,\\ 0, & otherwise.\end{cases} \tag{14}$$

For specific sample i, it is considered as "easy" if its losses smaller than age parameter τ, and the v_i will be assigned as 1; otherwise, it will be assigned as $v_i=0$. Samples with loss values of no more than age parameter τ will be selected for the training period.

Once V is measured, we enlarge the value of τ to include more samples with bigger losses in the training period. The iteration will stop until convergence. The whole algorithm for solving the SPS-NL Model is presented in Algorithm 1.

Algorithm 1.

Input: Training dataset $\{X_{n\times p}, y_n\}$, τ and a step size: ϖ
Output: Model parameter β
Step 1: Set $v_i{}^o=1$ ($i=1, 2, \ldots, n$) and τ.
Step 2: Update β^t based on Eq. (13) and the CDA.
Step 3: Update V^t based on Eq. (14).
Step 4: $\tau\leftarrow\varpi\tau$.
Step 5: Let $t\leftarrow t+1$, if $t<E$ then repeat Steps 2 – 4.

3.3. Time complexity analysis

The coordinate descent algorithm (CDA) is used to updates β in the Algorithm 1, The CDA algorithm is a powerful technique for deal with regularization model especially with high dimensional data, because its asymptotic time complexity is only O(mnp), where m is the number of iterations, n is the number of training sample size, p is the number of genes, and the values of m and n are generally small. When updates V in Algorithm 1, its asymptotic time complexity is O(n). The whole asymptotic time complexity for Algorithm 1 is O($(mnp+n)^*E$) = $O(mnpE)$, where E is the number of iterations and usually small (less than 110 in our experiment). Therefore, the final asymptotic time complexity for Algorithm 1 is O(mnp) which is similar to the standard CDA. It is noted that the Lasso, $L_{1/2}$, SCAD-L_2, Elastic-net and SCAD-Net methods are solved by CDA respectively in this paper, which implies that the computational time for solving the SPS-NL is similar to these approaches.

4. Results

4.1. Simulation

In this part, we perform a simulation study to evaluate the feature selection, and prediction capacity of the framework. Five sparse logistic model techniques: Lasso, $L_{1/2}$, SCAD-L_2, Elastic-net and SCAD-Net are compared in the experiment. Simulation scheme is similar to Li's work (Li & Li, 2008).

We simulated a network with 200 different transcription factors (TFs). Each TF in this test regulated 10 genes with add up to 2200 genes in the simulated network. The dependent variable y was assigned to a binary value 0 or 1, and was related with the first 4 TFs and their target genes.

We presented four scenarios in the simulation. For each scenario, we simulated 200 samples in which half of them for training and the other half for testing.

In scenario 1, there are two of the TFs and their target genes were positively related with the dependent variable and the rest of two TFs and their target genes were negatively related with the

Table 1
Gene selection result of the simulation study.

ρ		Lasso		$L_{1/2}$		SCAD-L_2		ElasticNet		SCAD-Net		SPS-NL	
		P	TP	P	TP	P	TP	P	TP	P	TP	P	TP
1	0.3	51.8	5.9	51.1	8.6	378.0	22.4	338.0	17.8	116.0	23.2	100.6	30.1
	0.6	57.0	10.4	54.3	17.1	436.4	32.2	371.5	29.6	116.0	22.2	107.1	32.6
	0.9	60.0	12.4	58.0	19.7	493.2	39.2	396.6	34.1	147.5	30.1	123.2	39.3
2	0.3	43.2	5.0	36.3	7.1	363.0	14.6	302.7	15.7	93.5	30.2	84.3	32.0
	0.6	49.7	7.6	41.6	9.2	276.5	26.5	336.5	22.6	124.8	25.3	103.9	36.6
	0.9	51.6	8.8	43.5	9.2	363.9	31.0	363.5	28.4	205.7	38.6	115.5	41.9
3	0.3	60.6	6.1	48.2	10.8	535.6	20.8	409.3	17.5	168.0	22.5	102.3	26.9
	0.6	59.4	8.2	42.8	11.0	354.4	21.5	410.6	23.5	180.3	26.4	157.7	38.0
	0.9	61.9	10.0	51.5	13.8	500.2	36.4	438.4	28.3	197.7	17.9	102.7	29.4
4	0.3	57.8	0.7	40.5	4.8	461.1	15.7	403.6	12.7	119.3	19.8	95.9	23.4
	0.6	59.6	7.8	47.0	6.4	572.0	24.6	416.0	20.9	154.5	26.1	118.5	26.4
	0.9	56.0	5.9	54.4	8.0	447.0	25.3	436.3	20.1	230.8	30.6	137.2	34.5

dependent variable:

$$\beta = \left(3, \underbrace{\tfrac{3}{\sqrt{10}}, ..., \tfrac{3}{\sqrt{10}}}_{10}, -3, \underbrace{\tfrac{-3}{\sqrt{10}}, ..., \tfrac{-3}{\sqrt{10}}}_{10}, 5, \underbrace{\tfrac{5}{\sqrt{10}}, ..., \tfrac{5}{\sqrt{10}}}_{10}, -5, \underbrace{\tfrac{-5}{\sqrt{10}}, ..., \tfrac{-5}{\sqrt{10}}}_{10}, 0, ..., 0 \right)$$

Expression levels for the 200 TFs were simulated using a standard normal distribution. Each TF and its target genes were jointly distributed as a bivariate normal with a correlation of ρ. We generated the dependent variable as $y = [\text{Prob}\ (y = 1|X;\ \beta) > 0.5]$. Then, 10% samples were randomly selected to given an extra noise 5ε, where $\epsilon \sim U(0, 1)$.

In scenario II, gene expression data were simulated as same as the scenario I except that a TF could be both an activator and repressor at the meantime. The coefficient vector was defined as:

$$\beta = \left(3, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \underbrace{\tfrac{3}{\sqrt{10}}, ..., \tfrac{3}{\sqrt{10}}}_{7}, -3, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \underbrace{\tfrac{-3}{\sqrt{10}}, ..., \tfrac{-3}{\sqrt{10}}}_{7}, 5, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \underbrace{\tfrac{5}{\sqrt{10}}, ..., \tfrac{5}{\sqrt{10}}}_{7}, -5, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \underbrace{\tfrac{-5}{\sqrt{10}}, ..., \tfrac{-5}{\sqrt{10}}}_{7}, 0, ..., 0 \right)$$

Scenario III was similar to scenario I except that we decreased the impact of the target genes on the dependent variable. Moreover, 20% samples were randomly selected and are given an extra noise 8ε.

$$\beta = \left(3, \underbrace{\tfrac{3}{10}, ..., \tfrac{3}{10}}_{10}, -3, \underbrace{\tfrac{-3}{10}, ..., \tfrac{-3}{10}}_{10}, 5, \underbrace{\tfrac{5}{10}, ..., \tfrac{5}{10}}_{10}, -5, \underbrace{\tfrac{-5}{10}, ..., \tfrac{-5}{10}}_{10}, 0, ..., 0 \right)$$

Table 2
Classification prediction result of the simulation study.

ρ		Lasso Accuracy	$L_{1/2}$	SCAD-L_2	ElasticNet	SCAD-Net	SPS-NL
1	0.3	83.9%	83.8%	83.4%	83.3%	85.4%	93.2%
	0.6	79.6%	81.4%	79.4%	79.3%	80.6%	90.8%
	0.9	78.5%	81.6%	80.7%	79.3%	81.2%	90.3%
2	0.3	86.9%	89.8%	89.1%	86.9%	88.8%	95.5%
	0.6	84.8%	87.2%	85.6%	84.9%	85.4%	94.3%
	0.9	82.7%	84.8%	82.3%	82.0%	84.8%	95.9%
3	0.3	81.2%	82.0%	77.5%	72.9%	76.5%	86.5%
	0.6	81.4%	79.7%	80.3%	73.4%	74.4%	85.1%
	0.9	78.6%	81.5%	80.0%	71.4%	82.5%	86.7%
4	0.3	81.8%	84.0%	83.5%	73.4%	77.8%	91.1%
	0.6	81.4%	82.1%	82.7%	72.0%	74.6%	86.9%
	0.9	81.4%	83.4%	83.2%	72.6%	80.1%	88.8%

Scenario IV was similar to scenario III except we allowing transcription factors to react to both activators and repressors.

$$\beta = \left(3, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \underbrace{\tfrac{3}{10}, ..., \tfrac{3}{10}}_{7}, -3, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \underbrace{\tfrac{-3}{10}, ..., \tfrac{-3}{10}}_{7}, 5, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \underbrace{\tfrac{5}{10}, ..., \tfrac{5}{10}}_{7}, -5, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \underbrace{\tfrac{-5}{10}, ..., \tfrac{-5}{10}}_{7}, 0, ..., 0 \right)$$

In practice, k ($k = 3$, 5 or 10)-fold cross-validation (CV) method is a popular approach to tune the parameter. Different k-fold CV schemes may yield very similar prediction results (Singh-Blom et al., 2013; Zeng, Liao, Liu, & Zou, 2017). Moreover, reducing the number of CV intervals from ten to three leads reducing the computation time of the algorithm by over half. Therefore, A 3-fold cross-validation on one dimension or multi-dimensions procedure was applied to the training data set to identify the optimal tuning parameter(s). Genes with a non-zero coefficient in the estimated model were recognized to be related to the clinical variable.

The correlation coefficient ρ of genes are set to 0.3, 0.6, 0.9 respectively. The simulation procedure was repeated 1000 times. We report the method's feature selection capacity with two indicators, P and TP respectively. The P indicates the number of the non-zero coefficient in the model, and the TP indicates the number of the true non-zero coefficient in the real model. We also computed the classification accuracy on the test data set. Results are summarized for each model in Tables 1 and 2.

As shown in Table 1, our method was much accurate in identifying true genes (TP) compared to the other algorithms. For example, when in scenario 2 with $\rho = 0.9$, the average TP selected by the SPS-NL method was 41.9, almost reach the whole true 44 genes. Our method also has good performance in

Table 3
The datasets used in this paper.

Datasets [GEO]	Platforms	NSCLC	Controls	Samples
GSE18842	HG-U133-Plus2	46	45	91
GSE19804	HG-U133-Plus2	60	60	120
GSE31547	HG-U133A	30	20	50
GSE32863	HumanWG-6 v3.0	58	58	116
GSE40419	HiSeq 2000	87	77	164
GSE10072	HG-U133A	58	49	107
GSE43458	HuGene-1.0-st	80	30	110
Total		**419**	**339**	**758**

classification prediction. As shown in Table 2, in all four scenarios, the proposed technique presented much higher precision compared to Lasso, $L_{1/2}$, SCAD-L_2, Elastic net and SCAD-Net logistic regression.

We also performed the simulation tests based on holdout and Leave-One-Out Cross Validation (LOOCV) methods. Results of gene selection and classification prediction by holdout and LOOCV were similar with the results under 3-fold CV, that the SPS-NL method is comparable or better efficient than the other methods.

There results indicate that the proposed method is an efficient prediction and gene selection tool to deal with the complex data such as small *n*, big *p*, highly correlated and noised.

4.2. Non-small cell lung cancer data analysis

4.2.1. Model development

To verify the proposed framework performance, we gathered and preprocessed several non-small cell lung cancer (NSCLC) gene expression datasets (Table 3).

Every dataset was mapped to a unique official gene symbol, and we summarized multiple probes sets that mapped to the same gene by their mean expression value. We used the subset of common genes represented among them, with a total of 10,881 remaining genes. After that, the training data (GSE18842, GSE19804, GSE31547, GSE40419) was homogenized with ComBat (a function in *sva*, R), and merged into a single large dataset with $N = 541$ samples includes 281 NSCLCs and 260 healthy controls. The validation data includes GSE10072 and GSE43458 respectively.

We downloaded the biological interaction network from the BioGrid (https://thebiogrid.org). The network is comprised of 14,355 genes or proteins and 324,663 interactions. By integrating the gene expression data with the prepared network, the final network L includes 10,881 genes and 207,934 edges.

Five approaches are compared with our proposed method: logistic regression with the Lasso, $L_{1/2}$, SCAD-L_2, Elastic-net and SCAD-Net. The optimal regularization parameters, or tuning parameters (which balance the tradeoff between data fit and model complexity), of the SPS-NL were tuned using 3-fold cross-validation (CV) on multi-dimensions in the training set ($N = 521$ samples). The final classifier model (with 90 genes and 1.97% training classification error) was built with the estimated tuning parameters using all the training data. The model's cut-off point was determined by the point that yielded the highest sum of sensitivity and specificity.

As demonstrated in Table 4, the SPS-NL approach provided the best training performance, with only 1.97% training error. This result is much better than competing approaches. For example, Lasso achieved 5.12% training error almost 2.6 times that of the method proposed here. Also, the SPS-NL method also outperforms the SCAD-Net method, which did not include an SPL mechanism, suggesting that the SPL strategy works well for combined gene expression data.

Table 4
Model training performances for each method.

Method	Training error	No. of selected genes
Lasso	5.12%	63
$L_{1/2}$	5.01%	49
SCAD-L_2	4.39%	96
Elastic net	4.64%	127
SCAD-Net	3.35%	138
SPS-NL	1.97%	90

Table 5
Performances on the validation sets.

Dataset	Method	Accuracy	AUC
GSE10072	Lasso	91.59%	92.32%
	$L_{1/2}$	92.52%	93.26%
	SCAD-L_2	94.39%	96.19%
	Elastic net	93.46%	94.11%
	SCAD-Net	94.39%	96.54%
GSE43458	Lasso	90%	91.74%
	$L_{1/2}$	93.64%	95.8%
	SCAD-L_2	93.64%	95.95%
	Elastic net	92.73%	95.61%
	SCAD-Net	94.45%	97.3%

4.2.2. Evaluation of validation sets

In this section, validation testing is performed on our NSCLC model. Classification accuracy and AUC under the receiver operating characteristic (ROC) analysis are reported.

We first applied our SPS-NL model to the clinical data GSE10072, the sample size $N = 107$ including NSCLC $n = 58$ and healthy controls $n = 49$. The AUC for the diagnostic test of NSCLC in this data was calculated to be 0.999 (Fig. 2-A1). We selected the threshold that achieved the highest sum of sensitivity and specificity in the data. The model achieved a sensitivity of 98.3%, a specificity of 100%, and an accuracy of 99.1% with only one misclassified sample. Moreover, the test scores were significantly different between cases and healthy controls ($p < 0.001$, *t*-test; Fig. 2-A2).

We then applied the model to the clinical data GSE43458, the sample size $N = 110$ including NSCLC $n = 80$ and healthy controls $n = 30$. The AUC for the diagnostic test of NSCLC in this data was determined to be 0.996 (Fig. 2-B1). Again, the selected cutoff point was the point that achieved the highest sum of sensitivity and specificity in the data. With the cutoff point selected, we observed a sensitivity of 97.5%, a specificity of 100%, and an accuracy of 98.2% with only two misclassified samples. We also observed a highly significant difference in the test score to be an NSCLC patient for cases compared with healthy controls ($p < 0.001$, *t*-test; Fig. 2-B2).

Table 5 shows competitor performance with the two validation datasets, for which the proposed method is very competitive. For example, the best accuracy result for the GSE10072 is 94.39% which is worse than our method (99.1%), with similar results from GSE43458.

This outcome indicates that the model built by the SPS-NL method from a large integration training dataset has a strong generalization capacity.

4.2.3. A brief biological discussion on signature genes by the SPS-NL

We observe that the selected genes are associated with NSCLC in many previous studies. For example, TP53 (Tumor protein p53) and KRAS (K-ras), play a critical role in cell proliferation, as well as cancer development and prognosis (Ling, Fabbri, & Calin, 2013). CLDN-5 (Claudin-5), helps regulate the rate of molecular movement in the intercellular space between the cells of an epithelium, and could be used as an additional diagnostic tool for

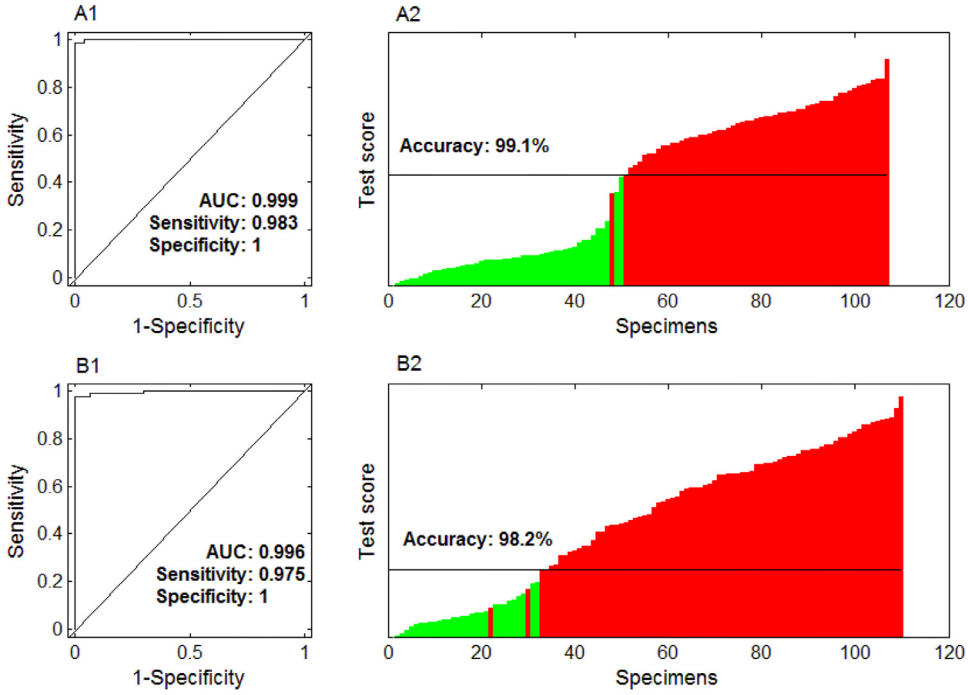

Fig. 2. Performance of the SPS-NL classifier when applied to the independent datasets GSE10072 and GSE43458. A1: receiver operating characteristic (ROC) curve analysis for the SPS-NL classifier applied to the GSE10072 (sample size $N = 107$ including NSCLC $n = 58$ and healthy controls $n = 49$); **A2:** test scores to be a case of all samples from GSE10072 were ranked. Healthy control cases are colored in green and NSCLC in red; **B1:** ROC curve analysis for the SPS-NL classifier applied to the GSE43458 (sample size $N = 110$ including NSCLC $n = 80$ and healthy controls $n = 30$); **B2:** test scores to be a case of all samples from GSE43458 were ordered. Red for NSCLC and green for healthy controls. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

adenocarcinoma as shown in the study by (Paschoud, Bongiovanni, Pache, & Citi, 2007). CA12 (Carbonic anhydrase 12) is overexpressed in various tumors, and it helps maintain extracellular acidic pH and may play a role in the cancer cell microenvironment by helping cancer cell reproduction and metastasis (Ivanov et al., 2001). SDC1 (Syndecan-1) is an omnipresent and essential extracellular matrix proteoglycan that affects basic fibroblast growth factor binding and activity. SDC1 is suggested to be involved in many major processes of tumorigenesis (Kim et al., 2015). MMP-12 (Matrix metalloproteinase-12) plays a key role in the decomposition of the extracellular matrix in cell reproduction, and tissue remodeling processes. MMP-12 expression is significantly correlated with NSCLC development and metastasis, and may be a valuable therapeutic target (Hofmann et al., 2005; Peng & Yang, 2017).

Fig. 3 shows the difference in the selected 90-signature gene expression between NSCLCs and healthy controls in primitive datasets GSE18842, GSE19804, GSE31547, GSE32863 and GSE40419 by using heat map analysis, respectively. For example, the expression of genes AGER and KARS is much higher in NSCLC patients than in healthy control patients. The expression of gene FABP4 is significantly lower in NSCLC patients than in healthy control patients. The heat map analysis for validation sets GSE10072 and GSE43458 is shown in Supplementary Figure S1, in which the similar differences were also observed. These results imply that the roles of these genes are more prominent in other genomic levels.

4.2.4. Network analysis

The 90 signature genes selected by SPS-NL were mapped to the BioGird network and analyzed using the Reactome pathways analysis (http://www.reactome.org). We reported the connected biological subnetworks in Fig. 4A. Subnetworks identified by SPS-NL allow investigators more easily to focus on key genes for subsequent downstream functional analyses. For example, the subnetwork connecting MEN1 and TP53, in which MEN1 encodes the nuclear protein menin, acts as a tumor suppressor in lung cancer and is often inactivated in human primary lung adenocarcinoma (Wu et al., 2012), implying that MEN1 and the other tumor suppressor TP53 may play a synergistic role in governing tumor activation and suppression in the development of lung cancer.

A subnetwork includes GATA1, GATA2, GATA3 and LMO2, and mainly relates to hematopoietic function. GATA1, GATA2 and GATA3 are expressed mainly in hematopoietic cell lineages (Vicente, Conchillo, García-Sánchez, & Odero, 2012). GATA1 is a critical transcription factor for the maturation of erythroid and megakaryocytic cells. GATA2 expression has a wide distribution among hematopoietic cells, with prominent expression in early progenitors, megakaryocytes and in mast cell lineages. GATA3 mainly influence the development of T lymphocytes. Lmo2 is a small protein composed of two LIM domains and is expressed in many tissues including hematopoietic precursors (Love, Warzecha, & Li, 2014). This hematopoietic-related subnetwork could be an important interplay in the development of NSCLC.

We summarized key pathways for the 90 signature genes (Fig. 4B) and all significant ($p<0.05$) pathways (Supplementary File 1). In total, these 90 signature genes are enriched in 49 distinct pathways. Some of the most significant biological processes include degradation of the extracellular matrix, syndecan interactions, collagen degradation and interleukin-4 and interleukin-13 signaling.

The extracellular matrix (ECM) serves various functions and is a critical component of the cellular microenvironment. ECM remodeling is a key mechanism for regulating cell differentiation, including processes such as branching morphogenesis, angiogenesis, bone remodeling, wound repair, the establishment and maintenance of stem cell niches (Lu, Takai, Weaver, & Werb, 2011). The core syndecan protein has three to five heparan sulfate or chondroitin sulfate chains, which interact with many ligands including

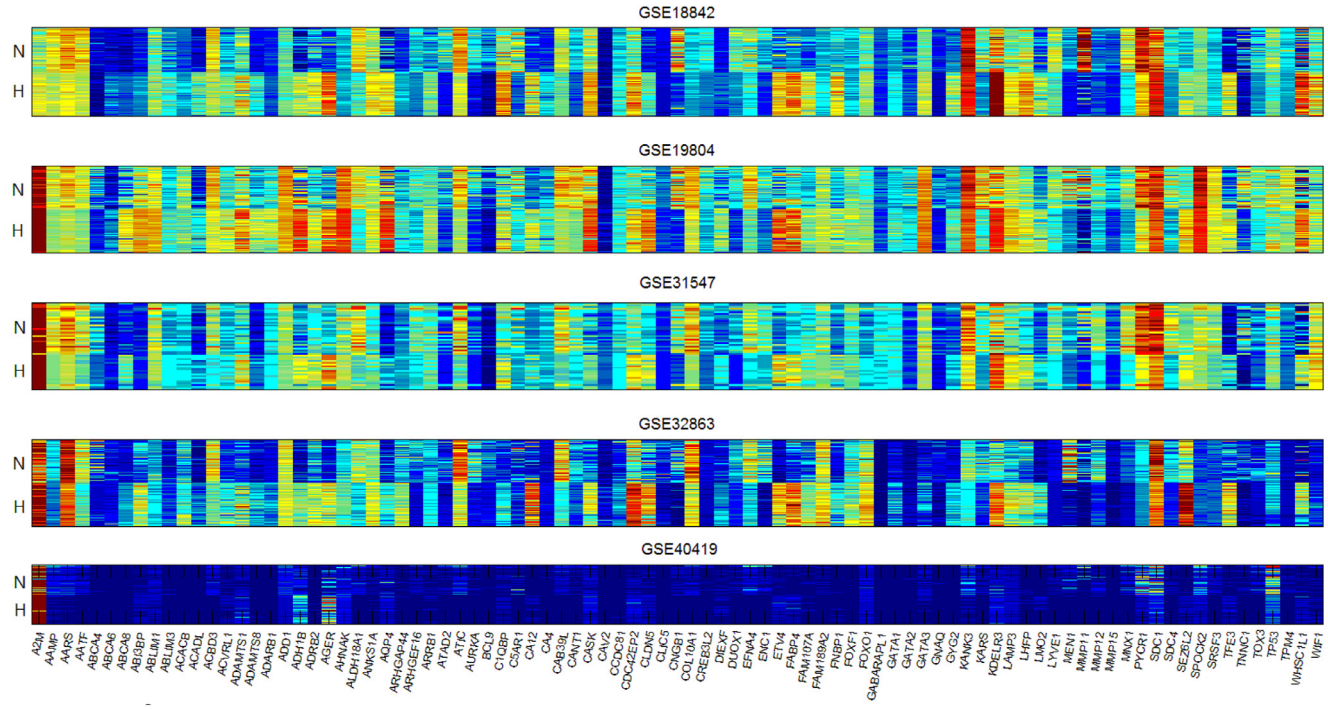

Fig. 3. The heat map of the gene signatures selected by SPS-NL approach for the original profiles. N: samples with NSCLC; H: samples with healthy controls. Red indicates high expression and blue indicates low expression. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

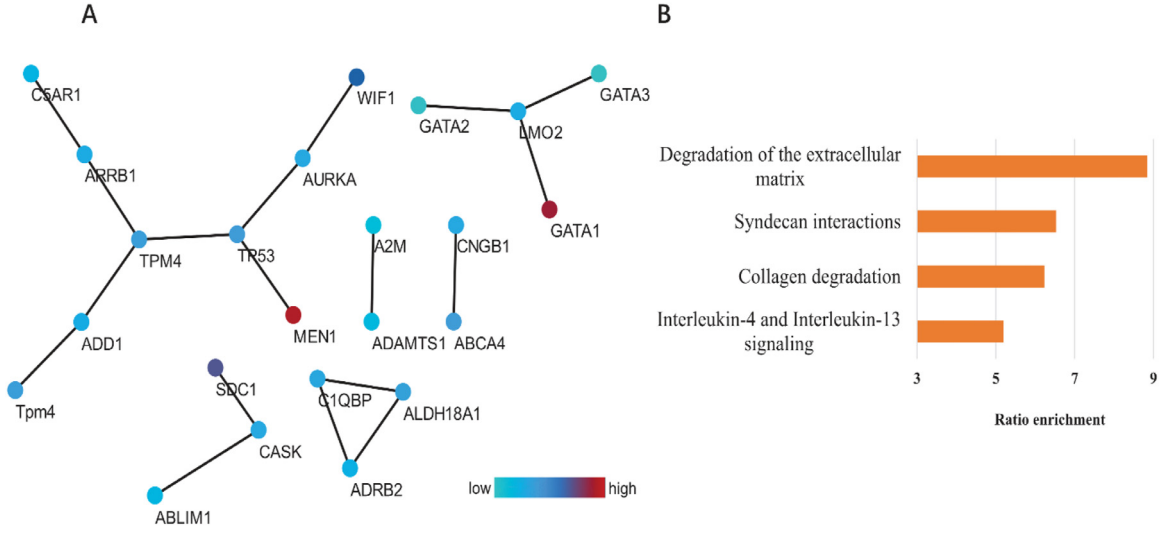

Fig. 4. The biological sub-network (A) and pathway analysis (B). Ratio enrichment indicates the functional significance of a gene module with -log(*p*-value).

fibroblast growth factors (FGFs), vascular endothelial growth factor (VEGF), transforming growth factor-beta (TGF-beta), fibronectin (FN), collagen, vitronectin (VTN) and several integrins. Such interactions may induce cancer cell proliferation of cancer cells (Alexopoulou, Multhaupt, & Couchman, 2007). Collagen is the key structural component of connective tissue and its degradation is a crucial mechanism in morphogenesis, tissue remodeling, and repair, and may reflect disease pathogenesis (Jabłońska-Trypuć, Matejczyk, & Rosochacki, 2016). Interleukin-4 (IL-4) is a typical pleiotropic T helper 2 cytokine involved in immunology during carcinogenesis. IL-4 causes G1-phasecell-cycle arrest of NSCLC cell lines expressing the interleukin 4 receptor, and can regulate modest to moderate antiproliferative activity *in vitro* and *in vivo* in animal models of human lung tumors (Essner, Huynh, Nguyen, Morton, & Hoon, 2000).

Cancer biomarkers have key potential benefits for patients, especially in contributing to personalized medicine and improved biomarkers should fundamentally lead to improvements in outcomes and more efficient, safe and cost-effective use of health resources (Ghosh, Begum, Sarkar, Chakraborty, & Maulik, 2019; Sayed, Nassef, Badr, & Farag, 2019; Zareizadeh, Helfroush, Rahideh, & Kazemi, 2018). Combining the results from Figs. 2-4, the gene-signatures selected by the SPS-NL have provided potential biomarkers in NSCLC. Moreover, the proposed method has identified the potential relationships in NSCLC with biologic significance. In a word, the proposed method has allowed the researcher to

more easily identifies a potential biomarker for functional studies or downstream applications.

5. Discussion and conclusion

In cancer genomic research with high-dimensional genetic measurements, the integrative analysis offers an effective solution of pooling information across various independent datasets and can lead to improved biomarker selection. In this paper, we have proposed a novel integrative analysis framework. In particular, the framework utilizes a new learning strategy (self-paced learning, SPL) and a novel feature selection method (SCAD-Net penalty). The traditional integrative analysis methods seek to combine multiple datasets into an integrated dataset and then analyze the data directly. However, such integration methods cannot eliminate internal bias, and may even add a new random noise and estimation errors to the integrated dataset, reducing the statistical power of the integrative analysis. Advancing from the published studies, we consider the SPL into our framework. This approach allows learning from the low-level noise samples first to develop a basic or universal biological knowledge, and when it is "powerful" enough, then learning the high-level noise samples to improve knowledge structure. Biomarker selection is a crucial part of the integrative analysis. The SCAD method is a popular biomarker selection method. However, the SCAD is proposed using purely computational points without any prior biological structure information. Advancing from the published studies, in this paper, we propose the SCAD-Net penalty to integrates the biological network knowledge. Moreover, some theoretical investigations of the SCAD-Net penalty are discussed. We combine the SPL with the SCAD-Net penalty, and couple with a logistic regression model to fulfill the integrative analysis framework (SPS-NL). We conduct a comprehensive simulation analysis, and an experiment on several large lung cancer datasets. The experimental results show that the proposed framework is promising. The proposed framework selects gene sets that are more coherent across datasets. Moreover, the selected genes have satisfactory stability and better prediction performance. Together, we have provided a new and efficient integrative method for biological research that help turns information from various gene-expression datasets into knowledge.

We use the hard regularization function $f(v;\ \tau) = -\tau v$ as the self-paced (SP) penalty. It may be promising to adopt the variants of this function to the proposed framework, *i.e.*, the linear function $f(v;\tau) = \tau(\frac{1}{2}v^2 - v)$ or the mixture function $f(v;\tau;\psi) = \frac{\psi^2}{v+\psi/\tau}$ (Meng et al., 2017). The SPS-NL method needs to tune across multiple penalized parameters, which are tuned by the grid search method with 3-fold cross-validation (CV) in this paper. Recently, evolutionary computations (EC) approaches have been used to tune the penalized parameter in the regularization problem for their global optimization capabilities (S. Wang, Shen, Chai, & Liang, 2019). The use of EC method instead of the CV approach to tune the penalized parameters in our SPS-NL model may improve the performance of the proposed framework. The logistic regression model involves the proposed framework for the lung cancer data. The applicability of the framework is relatively "independent" of the loss function. Therefore, alternative models, such as the Cox model or the AFT model, can be suggested. A more comprehensive examination of the framework with other cancer data/models will be pursued in the future. Limitations of this study also include a lack of more detailed analysis of the selected genes or pathways.

Acknowledgments

The authors are very much indebted to the anonymous reviewers, whose constructive comments are valuable for strengthening the presentation of this paper.

Funding

This research was supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences [18YJCZH054], National Natural Science Foundation of Guangdong [2018A030307033], Special Innovation Projects of Universities in Guangdong Province [2018KTSCX205], High-level Colleges Talent Project of Guangdong [2013–178], and Macau Science and Technology Development Funds [0055/2018/A2] of Macau SAR of China.

Conflicts of interest

None.

Authors' Contributions

Hai-Hui Huang developed the method, performed the analysis and drafted the article. Yong-Liang initiated, and supervised the project. All authors have read the final version of the article and have approved submission of the manuscript to "Expert Systems with Applications".

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2019.06.016.

References

Alexopoulou, A. N., Multhaupt, H. A. B., & Couchman, J. R. (2007). Syndecans in wound healing, inflammation and vascular biology. *The International Journal of Biochemistry & Cell Biology, 39*(3), 505–528.

Ali, H. R., Rueda, O. M., Chin, S.-F. F., Curtis, C., Dunning, M. J., & Aparicio, S. A. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology, 15*(8), 431.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (pp. 1–8). ACM Press.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., & Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics, 20*(1), 105–114.

Chen, J., Zhang, S., & C. , S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics, 32*(11), 1724–1732.

Dang, E., Yang, S., Song, C., Jiang, D., Li, Z., Fan, W., & Yang, K. (2018). BAP31, a newly defined cancer/testis antigen, regulates proliferation, migration, and invasion to promote cervical cancer progression. *Cell Death & Disease, 9*(8), 791.

Deng, K., Zhang, F., Tan, Q., Huang, Y., Song, W., Rong, Z., & Li, K. (2019). WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Analytica Chimica Acta, 1061*, 60–69.

Deshwar, A. G., & Morris, Q. (2014). PLIDA: Cross-platform gene expression normalization using perturbed topic models. *Bioinformatics, 30*(7), 956–961.

Essner, R., Huynh, Y., Nguyen, T., Morton, D. L., & Hoon, D. S. B. (2000). Functional interleukin 4 receptor and interleukin 2 receptor common γ-chain on human non–small cell lung cancers: Novel targets for immune therapy. *The Journal of Thoracic and Cardiovascular Surgery, 119*(1), 10–20.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., & Maulik, U. (2019). Recursive Memetic algorithm for gene selection in microarray data. *Expert Systems with Applications, 116*, 172–185.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., & Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology, 32*(1), 40–51.

Hofmann, H.-S., Hansen, G., Richter, G., Taege, C., Simm, A., & Silber, R.-E. (2005). Matrix metalloproteinase-12 expression correlates with local recurrence and metastatic disease in non-small cell lung cancer patients. *Clinical Cancer Research, 11*(3), 1086–1092.

Huang, H.-H., & Liang, Y. (2018). Hybrid L1/2 + 2 method for gene selection in the Cox proportional hazards model. *Computer Methods and Programs in Biomedicine, 164*, 65–73.

Huang, H.-H., Liang, Y., & Liu, X.-Y. (2015). Network-based logistic classification with an enhanced L 1 / 2 solver reveals biomarker and subnetwork signatures for diagnosing lung cancer. *BioMed Research International, 2015*, 713953.

Huang, H.-H., Liu, X.-Y., & Liang, Y. (2016). Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2 +2 regularization. *PloS One, 11*(5), e0149675.
Huang, H. H., Liu, X. Y., Li, H. M., & Liang, Y. (2017). Molecular pathway identification using a new L1/2 solver and biological network-constrained mode. *International Journal of Data Mining and Bioinformatics, 17*(3), 189.
Ivanov, S., Liao, S. Y., Ivanova, A., Danilkovitch-Miagkova, A., Tarasova, N., Weirich, G., & Stanbridge„ E. J. (2001). Expression of hypoxia-inducible cell-surface transmembrane carbonic anhydrases in human cancer. *The American Journal of Pathology, 158*(3), 905–919.
Jabłońska-Trypuć, A., Matejczyk, M., & Rosochacki, S. (2016). Matrix metalloproteinases (MMPs), the main extracellular matrix (ECM) enzymes in collagen degradation, as a target for anticancer drugs. *Journal of Enzyme Inhibition and Medicinal Chemistry, 31*(sup1), 177–183.
Jiang, L., Meng, D., Mitamura, T., & Hauptmann, A. G. (2014). Easy samples first: self-paced reranking for zero-example multimedia search. In *Proceedings of the ACM International Conference on Multimedia - MM '14* (pp. 547–556).
Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118–127.
Kim, S. Y., Choi, E. J., Yun, J. A., Jung, E. S., Oh, S. T., Kim, J. G., & Lee„ S. H. (2015). Syndecan-1 expression is associated with tumor size and egfr expression in colorectal carcinoma: a clinicopathological study of 230 cases. *International Journal of Medical Sciences, 12*(2), 92.
Kumar, M. P., Packer, B., & Koller, D. (2010). Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)* (pp. 1189–1197).
Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., & Nowe, A. (2013). Batch effect removal methods for microarray gene expression data integration: A survey. *Briefings in Bioinformatics, 14*(4), 469–490.
Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics, 24*(9), 1175–1182.
Liang, Y., Liu, C., Luan, X. Z., Leung, K. S., Chan, T. M., & Xu, Z. B. (2013). Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics, 14*, 198.
Ling, H., Fabbri, M., & Calin, G. A. (2013). MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature Reviews Drug Discovery, 12*(11), 847–865.
Love, P. E., Warzecha, C., & Li, L. (2014). Ldb1 complexes: The new master regulators of erythroid gene transcription. *Trends in Genetics : TIG, 30*(1), 1–9.
Lu, P., Takai, K., Weaver, V. M., & Werb, Z. (2011). Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harbor Perspectives in Biology, 3*(12), a005058.
Ma, S. S. (2009). Integrative analysis of cancer genomic data. *The 57th Session of the International Statistical Institute*, (33), 82–90.
Ma, Z., Liu, S., Meng, D., Zhang, Y., Lo, S. L., & Han, Z. (2018). On convergence properties of implicit self-paced objective. *Information Sciences, 462*, 132–140.
Meng, D., Zhao, Q., & Lu, J. (2017). A theoretical understanding of self-paced learning. *Information Sciences, 414*, 319–328.
Paschoud, S., Bongiovanni, M., Pache, J.-C., & Citi, S. (2007). Claudin-1 and claudin-5 expression patterns differentiate lung squamous cell carcinomas from adenocarcinomas. *Modern Pathology, 20*(9), 947–954 .
Peng, X., & Yang, Y. (2017). Algorithms for interval-valued fuzzy soft sets in stochastic multi-criteria decision making based on regret theory and prospect theory with combined weight. *Applied Soft Computing, 54*, 415–430.
Qi, L., Chen, L., Li, Y., Qin, Y., Pan, R., Zhao, W., & Guo, Z. (2016). Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: A case study for resected stage I non-small-cell lung cancer. *Briefings in Bioinformatics, 17*(2), 233–242.
Qin, L.-X., Huang, H.-C., & Begg, C. B. (2016). Cautionary note on using cross-validation for molecular classification. *Journal of Clinical Oncology, 34*(32), 3931–3938.
Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812–1823.
Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. *Expert Systems with Applications, 121*, 233–243.
Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., & Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics, 24*(9), 1154–1160.
Singh-Blom, U. M., Natarajan, N., Tewari, A., Woods, J. O., Dhillon, I. S., & Marcotte, E. M. (2013). Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS ONE, 8*(5), e58977.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.Series B (Methodological)*, 267–288.
Vicente, C., Conchillo, A., García-Sánchez, M. A., & Odero, M. D. (2012). The role of the GATA2 transcription factor in normal and malignant hematopoiesis. *Critical Reviews in Oncology/Hematology, 82*(1), 1–17.
Walker, E., Hernandez, A. V., & Kattan, M. W. (2008). Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine, 75*(6), 431–439.
Wang, R., Su, C., Wang, X., Fu, Q., Gao, X., Zhang, C., & Wei, M. (2018). Global gene expression analysis combined with a genomics approach for the identification of signal transduction networks involved in postnatal mouse myocardial proliferation and development. *International Journal of Molecular Medicine, 41*(1), 311–321.
Wang, S., Shen, H.-W., Chai, H., & Liang, Y. (2019). Complex harmonic regularization with differential evolution in a memetic framework for biomarker selection. *PLOS ONE, 14*(2), e0210786.
Wu, Y., Feng, Z.-J., Gao, S.-B., Matkar, S., Xu, B., Duan, H.-B., & Jin, G.-H. (2012). Interplay between Menin and K-Ras in Regulating Lung Adenocarcinoma. *Journal of Biological Chemistry, 287*(47), 40003–40011.
Zareizadeh, Z., Helfroush, M. S., Rahideh, A., & Kazemi, K. (2018). A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework. *Expert Systems with Applications, 113*, 301–314.
Zeng, L., & Xie, J. (2012). Group variable selection via SCAD- L 2. *Statistics, 48*(1), 49–66.
Zeng, X., Liao, Y., Liu, Y., & Zou, Q. (2017). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14*(3), 687–695.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.