Identifying Gene Signatures for Cancer Drug Repositioning Based on Sample Clustering

Fei Wang, Yulian Ding, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu

Abstract—Drug repositioning is an important approach for drug discovery. Computational drug repositioning approaches typically use a gene signature to represent a particular disease and connect the gene signature with drug perturbation profiles. Although disease samples, especially from cancer, may be heterogeneous, most existing methods consider them as a homogeneous set to identify differentially expressed genes (DEGs) for further determining a gene signature. As a result, some genes that should be in a gene signature may be averaged off. In this study, we propose a new framework to identify gene signatures for cancer drug repositioning based on sample clustering (GS4CDRSC). GS4CDRSC first groups samples into several clusters based on their gene expression profiles. Second, an existing method is applied to the samples in each cluster for generating a list of DEGs. Then a weighting approach is used to identify an intergrated gene signature from all the lists of DEGs. The integrated gene signature is used to connect with drug perturbation profiles in the Connectivity Map (CMap) database to generate a list of drug candidates. GS4CDRSC has been tested with several cancer datasets and existing methods. The computational results show that GS4CDRSC outperforms those methods without the sample clustering and weighting approaches in terms of both number and rate of predicted known drugs for specific cancers.

Index Terms—Drug repositioning, sample clustering, gene signature, gene expression, drug perturbation

1 Introduction

Traditionally, drug discovery industry is mainly about the screening of chemicals to obtain a small set of potential compounds [1]. However, further studies are needed to identify their therapeutic effects on a particular disease. After that, the satisfied compounds are moving forward to animal tests and clinical trials [2]. This whole complex process is so long and expensive that it takes 10-15 years and 0.8-1.5 billion US dollars to bring a drug from theory to product [3]. To reduce the time and cost of drug discoveries, researchers propose to find new usages for existing drugs, which have passed the evaluation of human safety [4]. Several successful drug repositioning studies have been published, including sildenafil for erectile dysfunction [5], thalidomide for severe erythema nodosum leprosum and retinoic acid for acute promyelocytic leukemia [6]. However, most of the successful examples of drug repositioning are from phenotypic drug screening and target-based methods [7], [8].

In recent years, the advances of high-throughput technologies, which produce a huge amount of transcriptome data, provide a great opportunity for studying drug repositioning. Based on the transcriptome data, several databases have been proposed for drug repositioning. Lamb *et al*. construct a Connectivity Map (CMap) database [9], [10]. In the database, there are 6,100 profiles in CMap build 2, each measuring the expression values of 22,283 genes of a cell line in a particular drug perturbation culture. The total number of drug perturbations is 1,309. To increase the scale of perturbations and keep the cost at a low level, the Library of Integrated Network-Based Cellular Signatures (LINCS) is developed [11]. The LINCS database only measures the expression values of 978 genes directly and all other gene expression values are estimated according to the measured values. About 19,811 small compound drug perturbations and 1,319,138 profiles are contained in the LINCS database.

After the construction of CMap and LINCS databases, several computational drug repositioning approaches have been proposed (e.g., [12],[13]). These approaches first identify a gene signature of a particular disease and then calculate the connection scores between the gene signature and the perturbation profiles in CMap database and/or LINCS database. The drugs with a connection score smaller than a threshold are identified as potential drugs for the disease, which are called drug candidates. Usually, among drug candidates, there are some drugs whose treatments for the particular disease are known, which are called known drugs. In general, the number of predicted known drugs can demonstrate the accuracy of the gene signature generated by the prediction method.

Many studies have been done to identify differentially expressed genes (DEGs), which are candidates of a gene signature. To identify DEGs, gene expression data, which collect gene expression levels in different tissue samples, are

- *Fei Wang and Yulian Ding are with the Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, Saskatchewan S7N 5A9, Canada. E-mail: {few266, yud146}@mail.usask.ca.*
- *Xiujuan Lei is with the School of Computer Science, Shaanxi Normal University, Xi'an 710119, China. E-mail: xjlei@snnu.edu.cn.*
- *Bo Liao is with the School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China. E-mail: dragonbw@163.com.*
- *Fang-Xiang Wu is with the Division of Biomedical Engineering, Department of Mechanical Engineering and Department of Computer Science, University of Saskatchewan, 57 Campus Drive, Saskatoon, Saskatchewan S7N 5A9, Canada, and also with the School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China. E-mail: faw341@usask.ca.*

Manuscript received 16 Apr. 2020; revised 16 July 2020; accepted 24 Aug. 2020. Date of publication 26 Aug. 2020; date of current version 1 Apr. 2022. (Corresponding author: Fang-Xiang Wu.)
Digital Object Identifier no. 10.1109/TCBB.2020.3019781

needed. The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [14] is one of the most comprehensive gene expression databases. Based on gene expression data, the fold-change thresholding methods are first used to identify DEGs (e.g., [15], [16]). Each gene has a fold-change ratio between normal tissue samples and disease tissue samples. The genes whose fold-change ratios are larger than a threshold are identified as DEGs.

However, the fold-change thresholding methods do not take variability into account or can not guarantee reproducibility [17]. Then the statistic methods are commonly used to identify DEGs, such as the T-test [18] and Wilcoxon test [19]. Additionally, based on the fact that disease-related proteins tend to have a larger number of interactions and more shared neighbors than non-disease proteins [20], genes can be mapped to protein-protein interaction (PPI) networks and use network methods to identify DEGs (e.g., [21], [22], [23]). In many of these studies, the disease tissue samples are treated as a homogeneous set to identify a gene signature. However, the samples from the same complex disease (e.g., cancer) are still heterogeneous as the complex diseases may have several subtypes. Therefore, treating all disease samples as a homogeneous set may average off the differences among the samples. As a result, DEGs or gene signatures generated by these methods are not good enough and thus their performance for drug repositioning is degraded.

In this study, we propose a new framework to identify gene signatures for cancer drug repositioning based on heterogeneous sample clustering (GS4CDRSC). GS4CDRSC first groups cancer samples into a number of clusters based on their gene expression profiles. Second, an existing method is applied to the samples in each cluster for generating a list of DEGs. In the lists of DEGs, a weighting approach is used to give each of the gene a new weight and sort them in descending order. Then the top genes are identified as gene signatures for drug repositioning. Finally, a CMap tool is applied to predict potential drugs from the integrated gene signature.

To evaluate its performance for drug repositioning, GS4CDRSC is combined with three existing approaches, while the k-means algorithm is employed to perform sample clustering. All the approaches are used to deal with tissue samples and identify a gene signature of a particular cancer. Then the gene signatures are used for drug repositioning and each gene signature obtains a list of drug candidates. To evaluate the accuracy of the gene signatures, the prediction rate of known drugs on the list of drug candidates have been calculated. Based on the known drugs, other predicted drugs on the list have potential for the same treatment. From the experiments we can see that with the proposed GS4CDRSC, higher prediction rates are generated, which means that GS4CDRSC can improve the performance of drug repositioning methods. Finally, we give a discussion about the predicted potential drugs.

2 Method

Typically, the computational drug repositioning approaches contain two main steps [24]: (1) Identifying DEGs based on a number of tumor tissue samples and normal tissue samples from the GEO database or the like, and further determining a gene signature of the specific cancer based on its DEGs; (2) Calculating the connection (or correlation) scores between drugs and gene signatures.

In drug repositioning, the approaches for identifying a gene signature play an important role. In most approaches, such as the fold-change thresholding approaches (e.g., [15], [16], [25]), statistic approaches (e.g., [18], [19]) and network approaches (e.g., [21], [22], [23]), all the samples from patients with the same clinical diagnosed diseases are treated as a homogeneous set. Therefore, the results of existing methods are not satisfied.

One of the reason is the heterogeneity of cancer samples, which contain several subtypes. The subtypes of a cancer are small groups that the cancer can be divided into, based on certain characteristics of the cancer cells. According to the studies of cancer cells in the past decades, different hierarchies of subtypes are proposed. Taking the lung cancer for instance, two main histological subtypes are non-small cell lung cancer (NSCLC, 85% of all lung cancers) and small cell lung cancer (SCLC, 15% of all lung cancers) [26]. There are three subtypes under the NSCLC, which are squamous cell lung carcinoma, adenocarcinomas and large cell carcinomas. Additionally, when looking into the hierarchy of genes and moleculars, some gene mutation based subtypes are proposed, such as epidermal growth factor receptor (EGFR)-mutation, kirsten rat sarcoma viral oncogene homolog (KRAS)-mutation and anaplastic lymphoma kinase (ALK)-mutation [27]. In clinics, some subtypes share similar treatments [27]. In our study, based on the gene expression values of patient samples, we aiming to improve the performance of drug repositioning methods based on sample clustering.

2.1 The GS4CDRSC Framework

In this study, we propose the GS4CDRSC framework for drug repositioning, which focuses on improving the identification of the gene signature of a specific cancer. The pipeline of GS4CDRSC is shown in Fig. 1. Specifically, a clustering algorithm is first used to divide the cancer samples into several clusters, each of which is expected to be homogeneous. Then the existing methods are employed to identify DEGs and generate a gene list for each sample cluster. In the list, a weighting approach is proposed to give each of the DEGs a new weight and sort the DEGs in descending order. Then the top M genes are identified as a DEG list. An integrated gene signature is determined over all the DEG lists from different clusters. The genes which appears in most of the DEG lists are utilized to construct the integrated gene signature. Finally, the integrated signature is used to query the CMap database and obtain drug candidates for the cancer under consideration. The detailed steps are illustrated in the following subsections.

2.2 The Sample Clustering

The sample clustering algorithm is used to produce a number of clusters that each cluster contains homogeneous samples. In our proposed GS4CDRSC framework, the k-means algorithm is used for this purpose although other clustering algorithms can be used at this step. In the k-means algorithm, the smaller the differences within a cluster, the better the results are [28].

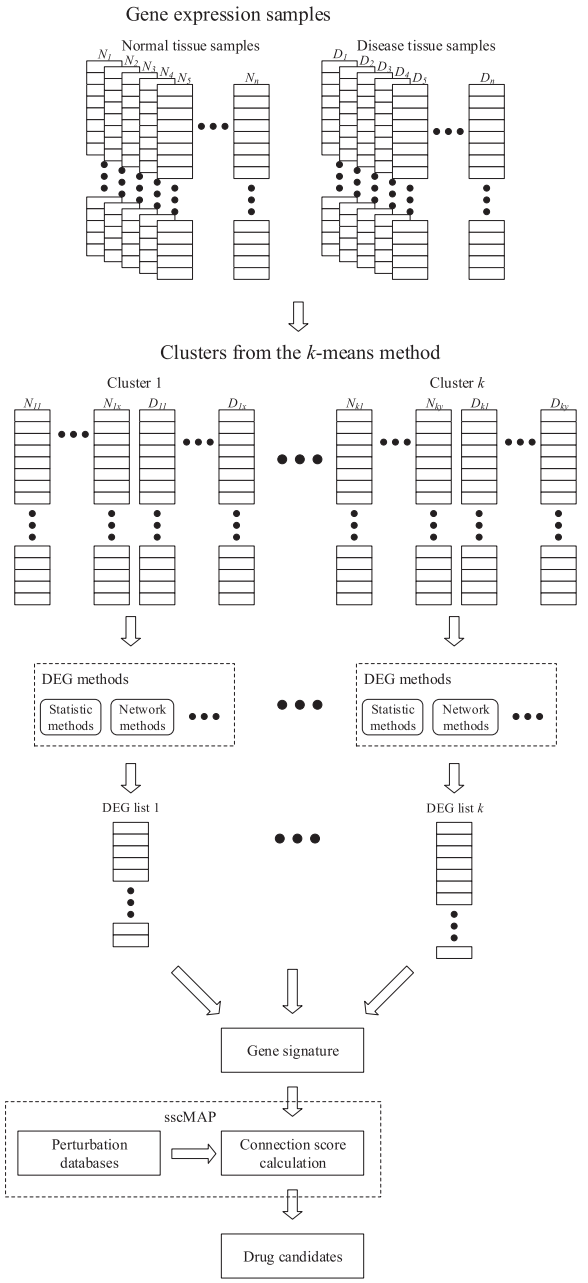

Fig. 1. The flowchart of the GS4CDRSC framework.

Given a set of samples $s = \{s_1, s_2, \ldots, s_n\}$, where each sample is a d-dimensional vector and d is the number of genes in a sample. The squared euclidean distance is used to measure the difference between two samples as follows:

$$dist(s_i, s_j) = \sum_{t=1}^{d} \big(s_i(t) - s_j(t)\big)^2. \quad (1)$$

The k-means algorithm is to obtain k clusters S= $\{S_1,S_2,...,S_k\}$ while the sum of distances within the clusters is the minimum. The objective of the k-means algorithm is to find the optimal S such that for a given k the following sum of squared errors (SSEs) is minimized:

$$J(S) = \sum_{i=1}^{k} \sum_{s \in S_i} dist(s, \mu_i), \quad (2)$$

where μ_i is the mean of the samples in cluster S_i. At the beginning of the algorithm, μ_i can be the profile of any sample. They are iteratively changed until the samples in the clusters are steady. As a result, all cancer samples are divided into k clusters. Then cancer samples in each subset and their corresponding normal samples are paired to make up a subset of samples for identifying DEGs and gene signatures in the following steps.

In GS4CDRSC, the k-means algorithm is based on DEGs and expected to obtain homogeneous subsets from all heterogeneous samples. Additionally, it is expected to reduce the effects of outliers in gene expression profiles. When measuring the values of genes in the microarray platforms, the accuracies of experiments are influenced by many factors, such as the pollution of the microarray, which produces few error values in some samples. When applying the k-means algorithm, the profiles with error values cannot affect all the clusters although they may affect some clusters, which improves its accuracy. Moreover, when considering the samples in a dataset as a whole set, some genes maybe averaged and ignored in the gene signature. The clustering algorithm is proposed to help to identify such genes. As shown in Table 4, most of the genes in the final signatures are totally new.

2.3 The DEG Identification for Each Subset

In GS4CDRSC, a list of DEGs is first generated from each subset of homogeneous samples which is obtained in Section 2.2. Then DEGs are used to identify gene signatures for drug repositioning. In this subsection, three DEG identification approaches are briefly described, including the moderated T test approach, the Wilcoxon test approach and a network based approach.

2.3.1 The Moderated T Test Approach

The T test is a pioneer approach to identify DEGs from gene expression profiles. However, the T test doesn't take into account the dependencies between genes. To address this weakness, the moderated T test is proposed [29]. Each gene is assigned a p-value based on its gene expression values across all samples. Meanwhile, a fold-change ratio is also assigned to the gene, according to its average expression value in normal tissue samples and that in tumor tissue samples. Then genes with small p-values and large fold-change ratios are identified as DEGs.

Suppose an expression value y_{gij} is from gene $g = \{1, \ldots, H\}$, array $i = \{1, \ldots, n\}$ and replicate $j = \{1, \ldots, m\}$. Let s_g^B be the between-array standard deviation, which is calculated as follows:

$$(s_g^B)^2 = \frac{m}{n-1} \sum_{i=1}^{n} (\overline{y}_{gi} - \overline{y}_g)^2, \quad (3)$$

where $\overline{y}_{gi}$ is the mean of the replicates of gene g on array i and $\overline{y}_g$ is the mean of gene g across all arrays. Let s_g^W be the within-array standard deviation, which is calculated as follows:

$$(s_g^W)^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{gij} - \overline{y}_{gi})^2. \quad (4)$$

Then a T score is calculated as follows:

$$T = \frac{\overline{y}_g \times \sqrt{nm[1+(m-1)\hat{\rho}]}}{s_g}. \tag{5}$$

where $\hat{\rho}$ is the correlation of gene between replicates and s_g is calculated as follows:

$$s_g^2 = \frac{\left\{\frac{(n-1)(s_g^B)^2}{1+(m-1)\hat{\rho}} + \frac{n(m-1)(s_g^W)^2}{1-\hat{\rho}}\right\}}{nm-1}. \tag{6}$$

A p-value is computed based on the T score. The False Discovery Rate (FDR) α is set to be 0.01 and is controlled by the Benjamini-Hochberg procedure [30] as follows:

$$p(M) \leq \frac{M}{H}\alpha, \tag{7}$$

where M is the length of gene signature and H is the number of genes in a sample. The largest M is set to be 100 to make sure that $p(M) \leq 1/H$. So that the maximum number of false genes in the signature is 1.

To construct a gene signature with M genes, the fold-change ratio between normal and tumor tissue samples are took into account. Let μ_1 and μ_2 be the average expression values of gene g in normal and tumor tissue samples, respectively. Then the fold-change ratio of gene g is $r = \mu_1/\mu_2$.

After generating the p-value and fold-change ratio of a gene, if its p-value is smaller than $1/H$ and its fold-change ratio is either larger than R or smaller than $1/R$, the gene is identified as a DEG candidate. Then the satisfied genes are sorted in ascending order based on their p-values. The ith gene in the list is given a weight of $(N-i+1)/N$, where N is the number of genes in the list. As a result, the top M genes are generated to identify a DEG list of the subset. Finally, k gene lists are obtained from k subsets.

2.3.2 The Wilcoxon Test Approach

In the Wilcoxon test approach, the p-value of a gene is based on a Z score [31]. Let the vector of differences between normal and tumor tissue samples be $d = \{d_1, d_2, \ldots, d_n\}$. Then the absolute values of differences are sorted in ascending order $D = \{D_1, D_2, \ldots, D_n\}$, and a sign vector $q = \{q_1, q_2, \ldots, q_n\}$ is associated with D, where D_i is the ith smallest absolute value in d. Let d_j be the corresponding value of D_i in d, if d_j is a positive value, then $q_i = 1$, otherwise $q_i = -1$. After that, a rank vector $v = \{v_1, v_2, \ldots, v_n\}$ is generated, where $v_i = i$. Particularly, if $D_i = D_{i+1} = \cdots = D_{i+j}$, the associated rank value is calculated as follows:

$$v_i = v_{i+1} = \cdots = v_{i+j} = \frac{\sum_{b=0}^{j}(i+b)}{j+1}. \tag{8}$$

Furthermore the Z score is calculated as follows:

$$Z = \frac{\left|\sum_{i=1}^{n} q_i v_i\right|}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \tag{9}$$

After that, the p-value is computed based on the Z score. The rest steps for obtaining DEGs are similar to those in Section 2.3.1.

2.3.3 The Network Based Approach

In the network based approach, one important step is to identify a DEG network from a PPI network [21]. In this study, we download the PPI data from BioGrid database [32]. Proteins in the PPI network and their corresponding genes in expression datasets are used to induce a gene network [33].

In the PPI network, we have some centrality measures that appropriate for it. The PPI networks have two properties: small world and free scale [34]. The bridging centrality works well in scale-free networks [35]. Jeong *et al.* propose that proteins with high the degree centrality are more likely to be associated with essential proteins [36]. Joy *et al.* conclude that the betweenness centrality is more likely to be essential than the degree centrality [37]. Closeness centrality and clustering coefficient are another commonly used topological parameters in biological network analyses [38], [39].

After obtaining a gene network from PPI network, DEGs generated from each cluster are mapped into the gene network. To obtain a DEG network for each cluster, DEGs and their direct neighbor genes in the gene network are retained. Then all other genes are deleted from the gene network. Finally, the gene network is transformed to a DEG network for each cluster. In the DEG network, the five centralities are used to measure the topological importance of genes, including the degree centrality, betweenness centrality, bridging centrality, closeness centrality and clustering coefficient.

Let the DEG network be $G = (V, E)$, where $V = \{v_1, \ldots, v_{n_1}\}$ is the set of n_1 vertices and $E = \{e_1, e_2, \ldots, e_{n_2}\}$ is the set of n_2 edges. The degree centrality of a vertex v is calculated as follows:

$$C_D(v) = d(v) = |N(v)|, \tag{10}$$

where $d(v)$ is the degree of vertex v, and $N(v)$ is the set of all neighbor vertices of v.

The betweenness centrality of a vertex v is calculated as follows:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{11}$$

where σ_{st} is the total number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the number of those shortest paths that pass through v.

The bridging centrality is calculated as follows:

$$C_R(v) = \frac{d(v)^{-1}}{\sum_{i \in N(v)} \frac{1}{d(i)}} \times C_B(v). \tag{12}$$

The closeness centrality of vertex v is calculated as follows:

$$C_C(v) = \frac{1}{\sum_s dis(v,s)}, \tag{13}$$

where $dis(v, s)$ is the distance between vertices v and s.

The clustering coefficient is calculated as follows:

$$C_L(v) = \frac{2 \times tri(v)}{|N(v)|(|N(v)|-1)}, \tag{14}$$

where $tri(v)$ is the number of triangles consisting of vertex v and its neighbors in G.

After generating all centralities of genes, the values are normalized to [0, 1] and each gene has a new weight in a centrality measure. The gene with a large centrality has a large weight. Then the five weights of a gene are summed up to a new weight. All genes are sorted in descending order according to the new weight. Then the ranked gene list is used to generate a gene signature.

2.4 The Gene Signature Determination

In previous sections, we have generated several gene lists from each of the methods applied in a dataset of the cancers. In this section, we describe a weighting approach to determine the gene signature from those gene lists.

Suppose we have L datasets of a cancer, each has k clusters. In each cluster, a number of tumor tissue samples and normal tissue samples are contained. One of three previous approaches is used to generate a gene list from a cluster. Then we are handling with $L \times k$ gene lists. Each gene in the list has a sign, either + or -, corresponding to the up-regulation or the down-regulation. To identify the up-regulation and the down-regulation, the average expression values of genes in tumor and normal tissue samples are calculated. An up-regulated gene has larger average expression value in tumor tissue samples than that in normal tissue samples. A down-regulated gene is the opposite, which has smaller average expression value in tumor tissue samples than that in normal tissue samples.

In addition, a gene on each of the list has a weight, which is based on three factors, including p-values, statistic powers and sample size. The p-values are used to describe the Type I error (also known as the false positive), while the statistical powers are used to describe the probability of Type II error (the false negative).

The first factor of weight depends on the rank of p-value in the gene list. The genes are sorted in ascending order based on their p-values. The ith gene in the list has the ith smallest p-value. Then its first factor of weight is $w_1 = (n_l + 1 - i)/n_l$, where n_l is the length of the gene list. If the gene is up-regulated, w_1 has a positive sign, and otherwise it has a negative sign. The larger the statistical power is, the lower probability Type II error occurs. The statistical power (SP) is the second part of weight, i.e., $w_2 = SP$. The sizes of the clusters are different. Then we use the size ratio $w_3 = n_c/n_d$ to be the third part of weight, where n_c and n_d are the sizes of a cluster and a dataset, respectively. Finally, the weight of a gene in a cluster is calculated as $w_1 \times w_2 \times w_3$. In this multiplication procedure, the normalization is not an essential step. The ranges of w_1 and w_2 are [0, 1] and the w_3 only has 2 possible values, when k is set to be 2. If we apply a normalization to the weights, the possible values of w_3 are 0 and 1, that the identified genes are totally based on the larger clusters. After the multiplication, the final weights of a gene are summed up on all $L \times k$ gene lists and sorted in descending order according to the absolute value. In this procedure, the normalization is not essential yet. Suppose the values of w_3 are different, then the ranges of $w_1 \times w_2 \times w_3$ in different clusters are not the same. If we apply a normalization, the w_3 fails to play a role. In addition, we think these three weight factors independently contribute to the final weight. According to the Bayesian rule, the multiplication of independently contribution is more reasonable. So we don't apply normalization after the multiplication.

After generating the final gene list, the top M genes are identified as the gene signature of the cancer. The largest value of M is described in Section 2.3.1. In the BH procedure, it tends to be a strong assumption that there are few signals. In the microarray studies, most genes are not related to the cancers [40]. After the multiplication of $w_1 \times w_2 \times w_3$, these assumptions are also satisfied, that more than 99.7% of the multiplied values in the experiments are 0.

2.5 The Connection Score Calculation

In this study, the sscMap platform [41] is used to calculate the connection score between a cancer (represented by its gene signature) and a drug candidate (represented by its induced cell line expression profile in the CMap database).

Given a cancer gene signature $G = \{g_1, g_2, \ldots, g_M\}$ and a drug induced profile $P_j (1 \leq j \leq N)$, where M is the number of genes in the integrated gene signature, and N is the number of drug induced profiles in the CMap database. The genes in P_j are sorted in descending order based on their expression values and $P_j(g_i)$ is denoted as the rank of gene g_i in the profile P_j. Then an intermediate connection score ICS is calculated as follows:

$$ICS(G, P_j) = \sum_{i=1}^{M} s(g_i)(I + 1 - P_j(g_i)), \tag{15}$$

where I is the number of genes in the drug induced profile.

A positive maximum connection score occurs when all the genes in a signature G are up-regulated genes and they are the same as the top s genes in a drug induced profile P_j. Then a positive maximum connection score $PMCS$ is calculated as follows:

$$PMCS(G, P_j) = \sum_{i=1}^{M} (I + 1 - i). \tag{16}$$

Then the connection score between a cancer gene signature G and a drug induced profile P_j is calculated as follows:

$$CS(g, P_j) = \frac{ICS(G, P_j)}{PMCS(G, P_j)}. \tag{17}$$

In general, the range of the connection score is $[-1,1]$. A connection score -1 indicates that the cancer gene signature and the drug-induced profile are most negative correlated, which is the best situation that the drug has a potential treatment for the cancer.

Additionally, a p-value is assigned to the connection score $CS(G, P_j)$. A number of random gene signatures are identified that the number of genes in a random gene signature is set to be n. Then the connection scores between the random gene signatures and the drug-induced profile P_j are obtained. After that, the p-value is the ratio of the random gene signatures whose connection scores are smaller than $CS(G, P_j)$. The p-value threshold is set to be $1/U$, where U is the number of drugs in the CMap database.

Finally, only the drugs whose connection scores are negative and p-values are smaller than $1/U$ are identified as drug candidates.

3 Experiments and Results

In this section, we apply our proposed GS4CDRSC framework on six cancers, including cervical cancer (CC), prostate cancer (PC), kidney cancer (KC), breast cancer, (BC) colorectal cancer (CRC), and non-small cell lung cancer (NSCLC). In the experiments, the gene signatures are generated by the three methods described in Section 2.3 with GS4CDRSC, including the clustering and weighting procedures. To make a comparison, the gene signatures are also generated by those methods without GS4CDRSC.

When evaluating the performance of drug repositioning methods, the prediction rate is proposed, which is the rate of the predicted known drugs to all the predicted drugs. The known drugs are the drugs that have shown their therapeutic effects in the particular cancer, alone or cooperate with other drugs. The annotations of all known drugs identified by GS4CDRSC are discussed in each case. For the approach to identify a gene signature of a cancer, the larger prediction rate with the gene signature can obtain, the better accuracy the gene signature should be. Then the other drug candidates have the potential to achieve the same treatments of the known drugs. We also discuss some annotations about the potential drugs.

3.1 Datasets

In this study, all gene expression datasets of the cancers are downloaded from the Gene Expression Omnibus (GEO) database [14]. In the GEO database, each cancer has several datasets. However, many of those datasets contain tumor tissue samples only. In our proposed framework, the generated datasets should contain both a number of cancer samples and a number of normal samples. The datasets of CC, PC, KC, BC, CRC and NSCLC are utilized in the experiments.

BC is the most common cancer in women, the cancer cells are formed in the breast. To study its gene signature, three gene expression datasets of breast cancer GSE10780, GSE15852 and GSE50948 are used in this study. PC is the most common cancer in men. It starts in the prostate. To study its gene signature, the dataset GSE46602 is used in this study. Lung cancer is the second most common cancer in both men and women. It is a disease in which the cancer cells form in the lung. About 85% of lung cancers are NSCLC. To study its signatures, three datasets GSE10072, GSE19804 and GSE27262 are used in this study. CRC is the third leading cause of cancer-related deaths in both men and women in the United States. The cancer cells form in the colon or rectum. To study its gene signatures, three datasets GSE21510, GSE41258 and GSE49355 are used in this study. CC is the fourth most common cancer in women. It is the cancer that starts in the cervix. It. To study its gene signature, the dataset GSE63514 is used in this study. KC is a disease that starts in the kidney. The terms kidney cancer and renal cell carcinoma (RCC) are often used interchangeably. To analyze its gene signature, the dataset GSE53757 is downloaded.

All the datasets are listed in Table 1 and belong to two platforms: GPL96 and GPL570. The GPL96 platform contains 22,283 probe sets, while the GPL570 platform contains 54,675 probe sets. Although the GPL 570 platform produces more information than the GPL96, the drug repositioning profiles in CMap are based on the GPL96 platform. To integrate the datasets from two platforms, we generate datasets with the 22,277 common probe sets among them. All the datasets are normalized using Robust multi-array average (RMA) method [42] and log2-transformed.

TABLE 1
The Number of Samples and Platforms in Each Dataset

Cases	Datasets	Platforms	Numbers of Samples
Breast	GSE50948	GPL570	80
	GSE15852	GPL96	86
	GSE10780	GPL570	84
Cervical	GSE63514	GPL570	48
Colon	GSE21510	GPL570	70
	GSE41258	GPL96	88
	GSE49355	GPL96	30
Kidney	GSE53757	GPL570	144
Lung	GSE10072	GPL96	48
	GSE19804	GPL570	96
	GSE27262	GPL570	50
Prostate	GSE46602	GPL570	28

In addition, we also study the associations of RNA_Seq datasets with the CMap database. The RNA_Seq datasets are downloaded from the Cancer Genome Atlas (TCGA) program in National Institutes of Health (NIH) [43]. To study the performance on RNA-Seq datasets, we generate 6 datasets from the database, including breast, bronchus and lung, cervix uteri, colon, kidney and prostate. In addition to the previous approaches, we utilize two new approaches to identify DEGs from RNA_Seq datasets, which are DESeq2 [44] and edgeR [45]. However, the prediction rates of the signatures generated from edgeR is 0 in all cases. Meanwhile, the prediction rates of DESeq2 is 0 in 3 cases, 0.1 in 2 cases and 0.2 in the colon tumor case. Then we try to scrutinize the possible reasons of it. The number of probes in RNA_Seq datasets is 60,483. Mapping the genes in the RNA_Seq dataset to the genes in CMap is an essential process. The Entrez gene IDs are used to be an intermediate to connect these two coding projects. However, only 13,845 probes in RNA_Seq data have their corresponding Entrez genes. Most of the information is lost, which leads to worse results. Thus we don't utilize the RNA_Seq data in the experiment.

3.2 Cluster Analysis

Before we compare the performance of the approaches with and without GS4CDRSC, we first determine the value of k for the k-means algorithm in GS4CDRSC. Actually, the determination of k for the k-means algorithm is a challenging issue. Although there is no best method for this issue in principle, one of the useful empirical methods is the Silhouette method [46]. In this study, the Silhouette method is utilized to generate validation of consistency within clusters.

As discussed in previous section, given a sample s_i in a cluster S_I, the mean distance between s_i and all other samples in cluster S_I is

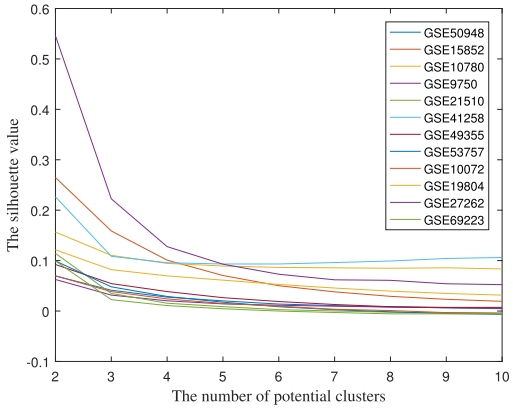

Fig. 2. The Silhouette values in each dataset. *k* is ranging from 2 to 10.

$$a(i) = \frac{1}{|S_I| - 1} \sum_{j \in S_I, i \neq j} dist(s_i, s_j). \tag{18}$$

Then the smallest distance between s_i and all samples in any other clusters is

$$b(i) = \min_{K \neq I} \frac{1}{|S_K|} \sum_{j \in S_K} dist(s_i, s_j). \tag{19}$$

Now we can calculate a silhouette value of a sample s_i:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, |S_I| > 1. \tag{20}$$

If $|S_I| = 1$, then $s(i) = 0$. Thus the average value of $s(i)$ over all samples is a measure of how appropriately the dataset have been clustered. A larger silhouette value refers to a better cluster result. As shown in Fig. 2, we have evaluate the 12 datasets in our experiments and the value *k* is ranging from 2 to 10. When *k* is set to be 2, the Silhouette values achieves the largest value in all the datasets. Then we utilized *k*=2 in our experiments.

3.3 Statistical Analysis

In our GS4CDRSC framework, we used the *k*-means algorithm to identify clusters from gene expression datasets. However, compared with the whole datasets, the size of each cluster is smaller. In this section, we learn the statistical influence of the changes of the sizes. The statistical power (SP) is the probability that it will reject a false null hypothesis.

Suppose we have n_p pairs of tumor-normal tissues. The average expression value of gene g in tumor tissues is m_t, while the standard deviation is sd_t. Then average expression value of gene g in normal tissue is m_n while the standard deviation is sd_n. The confidence level of the test is set to be 0.05, then the critical z score is 1.96 and -1.96. The SP is calculated as follows:

$$\begin{aligned} SP = \Phi\left(Z > 1.96 - \frac{(m_t - m_n)\sqrt{n_p}}{sd_n}\right) \\ + 1 - \Phi\left(Z > -1.96 - \frac{(m_t - m_n)\sqrt{n_p}}{sd_n}\right), \end{aligned} \tag{21}$$

where the z score has a corresponding confidence level. Then the SP of gene g is obtained.

TABLE 2
The Average Statistical Powers and the Number of Tumor-Normal Sample Pairs in all Clusters and Datasets

Cancers	Datasets	Cluster 1		Cluster 2		Undivided	
NSCLC	GSE10072	0.9873	15	0.9867	9	0.9873	24
	GSE19804	0.9892	23	0.5539	25	0.9878	48
	GSE27262	0.8153	9	0.8134	16	0.8153	25
CRC	GSE21510	0.9954	25	0.9950	10	0.9954	35
	GSE41258	0.9928	34	0.9903	10	0.8282	44
	GSE49355	0.8794	8	0.7011	7	0.7838	15
CC	GSE63514	0.8330	9	0.8199	15	0.3412	24
PC	GSE46602	0.8681	7	0.8624	7	0.8681	14
KC	GSE53757	0.9964	34	0.9967	38	0.9964	72
BC	GSE50948	0.9347	19	0.5966	21	0.8924	40
	GSE15852	0.9858	23	0.5680	20	0.9737	43
	GSE10780	0.9883	24	0.9887	18	0.9480	42

The SP is inversely related to the probability of making a Type II error. If a DEG has a large SP, it has a small possibility to be a non-DEG. To study the difference of SPs between the undivided dataset and the clusters, we calculate the SPs of all genes. To see the comparison of the changes clearly, we generate the average SPs of DEGs in each case and list them in Table 2. Among the 24 clusters, the average SP of 11 clusters are larger and those of 5 clusters are equal to those of the datasets. Then we can conclude that although the *k*-means algorithm decreases the size of profiles in each cluster, the SPs achieve benefits from it in a larger part.

3.4 Experiments

In the experiments, we applied our proposed GS4CDRSC framework to six types of cancers. To make a comparison between with and without using the clustering and weighting approaches, we utilized three different approaches to identify DEGs in our framework, as shown in Table 3. The prediction rates of all the comparisons are listed in Table 3. In most cases, we use the gene signature to identify 10 potential drugs. However, the numbers of potential drugs in some cases are less than 10.

In two cases CRC and BC, our GS4CDRSC framework achieves the higher prediction rates than without it. In the CC and KC cases, the approaches without our proposed framework cannot identify any known drug. In the PC and NSCLC cases, our GS4CDRSC framework could improve the prediction rates in two out of three approaches. The weakness of our proposed framework is that it cannot help to identify any drug of CC with moderated T test and KC with network-based approach. All the known drugs are discussed in Section 3.6.

3.5 Overlaps of the Signatures

In the experiments, one type of the comparison is the gene signatures between with and without the clustering and weighting approaches in our proposed framework. We generate the rates of overlapped DEGs among the signatures, as shown in the Table 4. In general, the rates are small. That most genes identified with our proposed framework are new. We also compared the numbers of overlapped DEGs

TABLE 3
The Prediction Rates by Two Types of Gene Signatures Identified in Six Cancer Cases and Three Approaches

Cancers	Approaches	Without	With
NSCLC	Moderated T test	0.20	0.20
	Wilcoxon	0.12	0.67*
	Network based	0.14	0.30
CRC	Moderated T test	0.20	0.40
	Wilcoxon	0.12	0.30
	Network based	0.12	0.60
CC	Moderated T test	0.00	0.00
	Wilcoxon	0.00	0.29*
	Network based	0.00	0.20
PC	Moderated T test	0.00	0.40
	Wilcoxon	0.25	0.25*
	Network based	0.00	0.60
KC	Moderated T test	0.00	1.00*
	Wilcoxon	0.00	0.25*
	Network based	0.00	0.00
BC	Moderated T test	0.13	0.30
	Wilcoxon	0.09	0.30
	Network based	0.07	0.20

**The number of drugs in the result is less than 10. Without: The signatures are generated from the datasets without our proposed framework. With: The signatures are generated from the GS4CDRSC framework with the clustering and weighting procedures.*

between two or three approaches with our proposed framework in Table 5.

3.6 Annotations of the Known Drugs

In this section, we discuss about the treatments of the known drugs in six cases. Many researchers have done a lot of studies about drugs and treatments. In all the cases, the histone deacetylase (HDAC) inhibitor is the largest type of drugs. Meanwhile, the HDAC inhibitors are used in the clinic of many cancers. Some drugs show individual treatment for a specific cancer. Some drug combinations are effective in some clinical trials.

3.6.1 Non-Small Cell Lung Cancer

Among the prediction results of our proposed framework, there are 4 drugs, clindamycin, glibenclamide, resveratrol and indomethacin, whose treatments have been studied. Glibenclamide is predicted by all three approaches, which is a medication used to treat diabetes mellitus type 2. It inhibit multidrug resistance protein 1 (MRP1) activities in human lung cancer cells and enhance the sensitivity of them to anticancer drugs [48]. Clindamycin is predicted by two of the approaches. It is a type of antibiotics. The combination of clindamycin and erlotinib is used for treating NSCLC and reducing the side effect of skin rash [47]. Resveratrol is a stilbenoid, a natural phytoalexin found in many food products, which can down-regulate the expression of survivin and induce the apoptosis in multidrug-resistant human NSCLC cells [49]. In addition, resveratrol can enhance the anti-tumor effects of the epidermal growth factor receptor (EGFR) inhibitor erlotinib in NSCLC cells [83]. Indomethacin is a nonsteroidal anti-inflammatory drug. It induces apoptosis in a doxorubicin-resistant lung cancer cell line through an MRP1-dependent mechanism [50].

TABLE 4
The Rates of the Overlapped Genes in the Signatures From the Approaches With Our Proposed Framework, Compared to the Approaches Without Our Proposed Framework

Cases	Moderated T test	Wilcoxon test	Network based
NSCLC	0.49	0.04	0.37
CRC	0.03	0.08	0.44
CC	0.18	0.05	0.18
PC	0.22	0.02	0.11
KC	0.03	0.00	0.52
BC	0.03	0.03	0.23

TABLE 5
The Rates of Overlapped Genes Between Two or Three Approaches in all Cases

Cases	Compare 1	Compare 2	Compare 3	Compare 4
NSCLC	0.22	0.40	0.21	0.11
CRC	0.13	0.23	0.18	0.06
CC	0.17	0.17	0.18	0.05
PC	0.16	0.15	0.05	0.05
KC	0.03	0.17	0.04	0.00
BC	0.16	0.13	0.09	0.02

All the approaches are combined with our proposed framework.
Compare 1: Between the moderated T test and network based approaches.
Compare 2: Between the moderated T test and Wilcoxon test approaches.
Compared 3: Between the network based and Wilcoxon test approaches.
Compared 4: Between all the 3 approaches.

3.6.2 Colorectal Cancer

Among the predicted drug lists, there are 12 drugs whose treatments have been studied, including tetrandrine, indomethacin, valproic acid, erastin, LY-294002, thioridazine, resveratrol, trichostatin A, methotrexate, trifluridine, etoposide and irinotecan. Valproic acid and trichostatin A are HDAC inhibitors. Valproic acid has been reported to impair the tumor-cell-induced angiogenesis [84]. It has also been shown to enhance the radiation response in CRC [53]. Trichostatin A reverses epithelialmesenchymal transition in colorectal cancer and induces apoptosis [57], [85].

Tetrandrine has anti-inflammatory, immunologic and anti-allergenic effects. It inhibits Wnt/β-catenin signaling and suppresses tumor growth of human colorectal cancer [51]. Indomethacin is a nonsteroidal anti-inflammatory drug. It suppresses growth of colon cancer via inhibition of angiogenesis in vivo [52]. Erastin is a small molecule capable of initiating ferroptotic cell death. It disrupts mitochondrial permeability transition pore (mPTP) and induces apoptotic death of colorectal cancer cells [54]. LY-294002 is a phosphatidylinositol 3-kinase (PI3K) inhibitor. It has been demonstrated to inhibit the cell growth and induce the cell apoptosis in colon cancer cell lines [55]. Thioridazine is an antipsychotic drug. It inhibits the proliferation of colorectal cancer stem cells through induction of apoptosis [56]. Resveratrol can depress the growth of colorectal aberrant crypt foci by affecting bax and p21 expression [58]. In further studies, it can inhibit the invasion and metastasis of CRC, in which long non-coding Metastasis Associated Lung

TABLE 6
The Known and Potential Drugs of Six Cancers Identified by the Three Approaches With GS4CDRSC

Cancers	Approaches	Known drugs in the results	Predicted potential drugs
NSCLC	Moderated T test	Clindamycin [47], Glibenclamide [48]	Clopamide, Ajmaline, Lobeline, Azacyclonol, Ampyrone, Danazol, Dirithromycin, Chlorzoxazone
	Wilcoxon	Resveratrol [49], Glibenclamide	Dirithromycin
	Network based	Indomethacin [50], Glibenclamide, Clindamycin	TTNPB, Anisomycin, Tetraethylenepentamine, Benzathine benzylpenicillin, Pirinixic acid, Lobeline, Ajmaline
CRC	Moderated T test	Tetrandrine [51], Indomethacin [52], Valproic acid [53], Erastin [54]	CP-320650-01, Mephenytoin, Beclometasone, Mycophenolic acid, Chlorhexidine, Oligomycin
	Wilcoxon	LY-294002 [55], Thioridazine [56], Trichostatin A [57]	Scopolamine, Zalcitabine, Pregnenolone, Fulvestrant, 6-Bromoindirubin-3′-oxime, 0297417-0002B, Maprotiline
	Network based	Resveratrol [58], Methotrexate [59], Trichostatin A, Trifluridine [60], Etoposide [61], Irinotecan [62]	0173570-0000, Hycanthone, Daunorubicin, PNU-0251126
CC	Moderated T test	NULL	NULL
	Wilcoxon	Sirolimus [63], LY-294002 [64]	Latamoxef, CP-645525-01, Zuclopenthixol, Picrotoxinin, Zalcitabine
	Network based	Sirolimus, Valproic acid [65]	0297417-0002B, SC-19220, CP-645525-01, Prochlorperazine, Oxantel, 15(S)-15-Methylprostaglandin E2, Adipiodone, Nortriptyline
PC	Moderated T test	Pyrvinium [66], Trichostatin A [57]	Prochlorperazine, Diclofenamide, Calmidazolium
	Wilcoxon	Geldanamycin [67]	0225151-0000, Dihydroergocristine, Tanespimycin
	Network based	Desipramine [68], Sirolimus [69], Withaferin A [70], Menadione [71], Thioridazine [72], Gossypol [73]	Thiostrepton, Isocarboxazid, 6-Benzylaminopurine, 0175029-0000
KC	Moderated T test	LY-294002 [74]	Irinotecan
	Wilcoxon	Anisomycin [75]	Fulvestrant, CP-690334-01, BCB000039
	Network based	NULL	Ciclopirox, Estropipate, Ethisterone, Letrozole, Etiocholanolone, Erastin, Benzathine Benzylpenicillin, Metergoline, Selegiline, Rifampicin
BC	Moderated T test	Metformin [76], Oligomycin [77], Danazol [78]	Primidone, Rilmenidine, Propidium iodide, Ozagrel, Oxybenzone, Iohexol, Merbromin, Chlorzoxazone
	Wilcoxon	Rosiglitazone [79], MS-275 [80], TTNPB [81]	Monorden, Indomethacin, Lasalocid, Iloprost, Nadolol
	Network based	Fulvestrant [82], Metformin [76]	Clopamide, Iloprost, Chlorzoxazone, Dicycloverine, Fludrocortisone, Dirithromycin

NULL in the table indicates that there's no result in the experiment.

Adenocarcinoma Transcript 1 (RNA-MALAT1) plays an important role [86].

Methotrexate is an immune system suppressant that also has antitumor treatments in breast cancer and lung cancer. The combination of leucovorin and fluorouracil with it is an active regimen in advanced colorectal cancer [59]. Trifluridine is an anti-herpesvirus antiviral drug. It has recently been approved for the treatment of adult patients with metastatic colorectal cancer [60]. Etoposide a chemotherapy medication used for the treatments of a number of types of cancer. It has anti-proliferative effects in colon cancer cells [61]. Irinotecan a medication used to treat colon cancer and small cell lung cancer. The treatment of it plus fluorouracil and leucovorin is better than a widely used therapeutic regimen of fluorouracil and leucovorin [62]

3.6.3 Cervical Cancer

In the results, only 3 drugs have been studied for their treatment of cervical cancer. LY-294002 is s potent inhibitor of numerous proteins, and a strong inhibitor of phosphoinositide 3-kinases (PI3Ks). Its PI3K inhibition produces significant radiosensitization, and increase apoptosis in human cervical cancer cell lines [64]. Sirolimus, also known as rapamycin, is a macrolide compound. It can significantly enhance the sensitivity of CaSki cells (a type of human cervical cancer cell lines) to paclitaxel, which is effectively against cervical cancer [63]. Valproic acid is used to treat certain types of seizures. It has shown its antitumor effects in NSCLC and CRC. It inhibits the in vitro angiogenic potential of human cervical cancer cells [65].

3.6.4 Prostate Cancer

There are 9 known drugs in the predicted results. Trichostatin A is an HDAC inhibitors and have shown antitumor effects in different types of cancers. It reduces cell invasion and migration abilities in prostate cancer cells [57]. Pyrvinium is a known drug for cervical cancer. Androgen receptor (AR) is a type of nuclear receptor. It has a key role in prostate cancer progression [87]. Pyrvinium can suppress prostate cancer cells through of endogenous AR in human prostate cancer cell lines [66], [88]. Gossypol is a nature phenol derived from the cotton plant. It is currently in phase II clinical trials as an adjuvant therapy for human prostate cancer [73]. Geldanamycin is an antitumor antibiotic that has inhibition of angiogenesis in prostate cancer cells [67].

Desipramine is a tricyclic antidepressant (TCA) used in the treatment of depression. It causes apoptosis via inducing

c-Jun NH2-terminal kinase (JNK)-associated caspase-3 activation[68]. Sirolimus shows treatment in both androgen dependent and independent prostate cancer cells [69]. Withaferin A is a steroidal lactone. It induces mitotic catastrophe and growth arrest in prostate cancer cells [70]. Menadione is an organic compound. The combination of ascorbate and menadione induces cell death in human prostate cancer cells [71]. Thioridazine have shown treatment in colorectal cancer. It significantly inhibited the growth of prostate cancer cells in vitro (including androgenindependent colonies) [72].

3.6.5 Kidney Cancer

Only 2 drugs have shown their treatment for kidney cancer. LY-294002 is a PI3K inhibitor and PI3KAkt signaling cascade is, in theory, an ideal therapeutic target for this kidney cancer [89]. The combination of LY-294002 with gefitinib suppresses the viability of gefitinib-resistant kidney cancer cell lines [74]. Anisomycin is an antibiotic which inhibits eukaryotic protein synthesis. It sensitizes human kidney cancer cells to the tumor necrosis factor (TNF)-related apoptosis-inducing ligand (TRAIL)-induced apoptosis [75].

3.6.6 Breast Cancer

There are 7 known drugs in the predicted results. Metformin and MS-275 have shown antitumor effects in a variety of cancers. Metformin is an AMP kinase-dependent growth inhibitor for breast cancer cells [76]. MS-275 is an HDAC inhibitor, that it inhibits the tumor progression, angiogenesis, and metastasis of breast cancer [80]. Oligomycin is a macrolide created by Streptomyces. It abolishes the growth of human breast cancer cells at remarkably low concentrations [77]. Danazol a medication used in the treatment of endometriosis. It is an effective treatment for advanced breast cancer [78]. Rosiglitazone is an antidiabetic drug. It sensitizes breast cancer cells to anti-tumor effects of TNF-α, CH11 and CYC202 [79]. Fulvestrant is a medication that used to treat hormone receptor (HR)-positive metastatic breast cancer [82]. Arotinoid acid (TTNPB) proves to be 100 times more effective than all-trans-retinoic acid (atRA), which also has great growth inhibition of breast cancer cells [81].

3.6.7 Discussions About the Predicted Drugs

In the experiments, we have identified some small compound drugs that have shown treatments against cancers and some drugs that may have potential treatments. In former subsections, we have talked about the treatments of the known drugs, which are side witnesses of the predicted drugs. In this section, we discuss about some of the predicted drugs that have anti-tumor effects in a variety of cancers.

Among the predicted results of NSCLC, danazol and TTNPB have shown some treatments against a variety of cancers, which denotes the potential anti-tumor effects on NSCLC. In the predicted drugs of CRC, chlorhexidine, daunorubicin and oligomycin are known drugs for different cancers. In the third CC case, nortriptyline has shown treatments on a many types of cancers. In the predicted drugs of PC, tanespimycin and thiostrepton are identified as anti-tumor agents in a variety of cancers. In the results of KC, irinotecan, fulvestrant and erastin have some treatments for different cancers. In the predicted drugs of BC, clindamycin, estradiol, gabexate and altretamine are anti-tumor agents in many cancers. Especially, altretamine is predicted by all three approaches with our proposed framework.

4 Conclusion

In this study, we have proposed a GS4CDRSC framework to identify a gene signature of a particular cancer for drug repositioning. After sample clustering, the existing DEG approach is performed a number of times based on the k clusters. At each time, a list of DEGs is identified from each cluster. Then the DEGs from all clusters are used to generate an integrated gene signature. Comprehensive experiments have been conducted to evaluate the performance of the proposed framework. The results demonstrate the effectiveness of GS4CDRSC in identifying a gene signature. With the proposed framework, the gene signatures identified from existing approaches can obtain more known drugs and the prediction rates of known drugs in drug candidates are larger than the approaches without the framework. In future, we would study more data and expand the applications of the proposed framework for drug repositioning.

Acknowledgments

This work was supported in part by the Natural Science and Engineering Research Council of Canada (NSERC), China Scholarship Council (CSC) and by the National Natural Science Foundation of China under Grant No. 61772552 and No. 61428209.

References

[1] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.
[2] S. M. Paul *et al.*, "How to improve R&D productivity: The pharmaceutical industry's grand challenge," *Nat. Rev. Drug Discov.*, vol. 9, no. 3, pp. 203–214, 2010.
[3] F. Emmert-Streib , S. Tripathi, R. M. Simoes, A. F. Hawwa, and M. Dehmer, "The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes," *Syst. Biomed.*, vol. 1, no. 1, pp. 20–28, 2013.
[4] M. L. Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, and J. R. Green, "A review of network-based approaches to drug repositioning," *Briefings Bioinf.*, vol. 19, no. 5, pp. 878–892, 2018.
[5] M. Boolell *et al.*, "Sildenafil: An orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction," *Int. J. Impotence Res.*, vol. 8, no. 2, pp. 47–52, 1996.
[6] J. Aronson, "Old drugs–new uses," *Brit. J. Clin. Pharmacol.*, vol. 64, no. 5, pp. 563–565, 2007.
[7] Z. Wu, Y. Wang, and L. Chen, "Network-based drug repositioning," *Mol. Biosyst.*, vol. 9, no. 6, pp. 1268–1281, 2013.
[8] F. Wang, X. Lei, B. Liao, and F. X. Wu, "Human protein complex signatures for drug repositioning," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform.*, 2019, pp. 42–50.
[9] J. Lamb *et al.*, "The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.
[10] J. Lamb, "The connectivity map: A new tool for biomedical research," *Nat. Rev. Cancer*, vol. 7, no. 1, pp. 54–60, 2007.
[11] A. B. Keenan *et al.*, "The library of integrated network-based cellular signatures NIH program: System-level cataloging of human cells response to perturbations," *Cell Syst.*, vol. 6, pp. 13–24, 2017.
[12] X. Zhou, M. Wang, I. Katsyv, H. Irie, and B. Zhang, "EMUDRA: Ensemble of multiple drug repositioning approaches to improve prediction accuracy," *Bioinformatics*, vol. 34, pp. 3151–3159, 2018.
[13] A. Peyvandipour, N. Saberian, A. Shafi, M. Donato, S. Draghici, and A. Valencia, "A novel computational approach for drug repurposing using systems biology," *Bioinformatics*, vol. 34, no. 16, pp. 2817–2825, 2018.

[14] T. Barrett *et al.*, "NCBI GEO: Archive for functional genomics data setsupdate," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, 2012.
[15] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes," *Proc. Nat. Acad. Sci. USA*, vol. 93, no. 20, pp. 10 614–10 619, 1996.
[16] J. DeRisi *et al.*, "Use of a CDNA microarray to analyse gene expression," *Nat. Genet.*, vol. 14, pp. 457–460, 1996.
[17] D. J. McCarthy and G. K. Smyth, "Testing significance relative to a fold-change threshold is a treat," *Bioinformatics*, vol. 25, no. 6, pp. 765–771, 2009.
[18] Q. Wen *et al.*, "Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies," *BMC Syst. Biol.*, vol. 9, no. 5, 2015, Art. no. S4.
[19] P. G. O'Reilly *et al.*, "Quadratic: Scalable gene expression connectivity mapping for repurposing FDA-approved therapeutics," *BMC Bioinf.*, vol. 17, no. 1, 2016, Art. no. 198.
[20] J. Xu and Y. Li, "Discovering disease-genes by topological features in human proteinprotein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.
[21] C. H. Huang, P. M. H. Chang, C. W. Hsu, C. Y. F. Huang, and K. L. Ng, "Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory," in *BMC Bioinf.*, vol. 17, no. 1, 2016, Art. no. S2.
[22] E. March-Vila *et al.*, "On the integration of in silico drug design methods for drug repurposing," *Front. Pharmacol.*, vol. 8, 2017, Art. no. 298.
[23] X. Liu, X. Chang, R. Liu, X. Yu, L. Chen, and K. Aihara, "Quantifying critical states of complex diseases using single-sample dynamic network biomarkers," *PLoS Comput. Biol.*, vol. 13, no. 7, 2017, Art. no. e1005633.
[24] F. Wang, X. Lei, and F. X. Wu, "A review of drug repositioning based chemical-induced cell line expression data," 2020. [Online]. Available: https://doi.org/10.2174/0929867325666181101115801
[25] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory, "RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis," *Bioinformatics*, vol. 22, no. 22, pp. 2825–2827, 2006.
[26] K. Inamura, "Lung cancer: Understanding its molecular pathology and the 2015 who classification," *Front. Oncol.*, vol. 7, 2017, Art. no. 193.
[27] L. West *et al.*, "A novel classification of lung cancer into molecular subtypes," *PLoS One*, vol. 7, no. 2, 2012, Art. no. e31906.
[28] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probability*, 1967, vol. 1, no. 14, pp. 281–297.
[29] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.
[30] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc.: Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.
[31] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics*, vol. 18, no. 11, pp. 1454–1461, 2002.
[32] A. Chatr-Aryamontri *et al.*, "The biogrid interaction database: 2015 update," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D470–D478, 2014.
[33] W. Lan, J. Wang, M. Li, W. Peng, and F. X. Wu, "Computational approaches for prioritizing candidate disease genes based on PPI networks," *Tsinghua Sci. Technol.*, vol. 20, no. 5, pp. 500–512, 2015.
[34] X. Lei, J. Tian, L. Ge, and A. Zhang, "The clustering model and algorithm of PPI network based on propagating mechanism of artificial bee colony," *Inf. Sci.*, vol. 247, pp. 21–39, 2013.
[35] W. Hwang, Y. R. Cho, A. Zhang, and M. Ramanathan, "Bridging centrality: Identifying bridging nodes in scale-free networks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 20–23.
[36] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
[37] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *J. Biomed. Biotechnol.*, vol. 2005, no. 2, p. 96–103, 2005.
[38] M. Jalili *et al.*, "Evolution of centrality measurements for the detection of essential proteins in biological networks," *Front. Physiol.*, vol. 7, 2016, Art. no. 375.
[39] C. C. Friedel and R. Zimmer, "Inferring topology from clustering coefficients in protein-protein interaction networks," *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. 519.
[40] J. X. Hu, H. Zhao, and H. H. Zhou, "False discovery rate control with groups," *J. Amer. Stat. Assoc.*, vol. 105, no. 491, pp. 1215–1227, 2010.
[41] S. D. Zhang and T. W. Gant, "sscMap: An extensible java application for connecting small-molecule drugs using gene-expression signatures," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. 236.
[42] R. A. Irizarry *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.
[43] R. L. Grossman *et al.*, "Toward a shared vision for cancer genomic data," *N. Engl. J. Med.*, vol. 375, no. 12, pp. 1109–1112, 2016.
[44] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DEseq2," *Genome Biol.*, vol. 15, no. 12, 2014, Art. no. 550.
[45] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
[46] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[47] P. Bidoli *et al.*, "Isotretinoin plus clindamycin seem highly effective against severe erlotinib-induced skin rash in advanced non-small cell lung cancer," *J. Thoracic Oncol.*, vol. 5, no. 10, pp. 1662–1663, 2010.
[48] L. Payen, L. Delugin, A. Courtois, Y. Trinquart, A. Guillouzo, and O. Fardel, "The sulphonylurea glibenclamide inhibits multidrug resistance protein (MRP1) activity in human lung cancer cells," *Brit. J. Pharmacol.*, vol. 132, no. 3, pp. 778–784, 2001.
[49] W. Zhao, P. Bao, H. Qi, and H. You, "Resveratrol down-regulates survivin and induces apoptosis in human multidrug-resistant SPC-A-1/CDDP cells," *Oncol. Rep.*, vol. 23, no. 1, pp. 279–286, 2010.
[50] D. J. A. De Groot , M. Van Der Deen , T. K. P. Le, A. Regeling, S. De Jong , and E. G. E. De Vries , "Indomethacin induces apoptosis via a MRP1-dependent mechanism in doxorubicin-resistant small-cell lung cancer cells overexpressing MRP1," *Brit. J. Cancer*, vol. 97, no. 8, pp. 1077–1083, 2007.
[51] B. C. He *et al.*, "Tetrandrine inhibits Wnt/β-catenin signaling and suppresses tumor growth of human colorectal cancer," *Mol. Pharmacol.*, vol. 79, no. 2, pp. 211–219, 2011.
[52] H. M. Wang and G. Y. Zhang, "Indomethacin suppresses growth of colon cancer via inhibition of angiogenesis in vivo," *World J. Gastroenterol.: WJG*, vol. 11, no. 3, p. 340–343, 2005.
[53] X. Chen, P. Wong, E. Radany, and J. Y. C. Wong, "HDAC inhibitor, valproic acid, induces p53-dependent radiosensitization of colon cancer cells," *Cancer Biother. Radiopharm.*, vol. 24, no. 6, pp. 689–699, 2009.
[54] H. Huo, Z. Zhou, J. Qin, W. Liu, B. Wang, and Y. Gu, "Erastin disrupts mitochondrial permeability transition pore (mPTP) and induces apoptotic death of colorectal cancer cells," *PLoS One*, vol. 11, no. 5, 2016, Art. no. e0154605.
[55] S. Semba, N. Itoh, M. Ito, M. Harada, and M. Yamakawa, "The in vitro and in vivo effects of 2-(4-morpholinyl)-8-phenyl-chromone (LY294002), a specific inhibitor of phosphatidylinositol 3-kinase, in human colon cancer cells," *Clin. Cancer Res.*, vol. 8, no. 6, pp. 1957–1963, 2002.
[56] C. Zhang, P. Gong, P. Liu, N. Zhou, Y. Zhou, and Y. Wang, "Thioridazine elicits potent antitumor effects in colorectal cancer stem cells," *Oncol. Rep.*, vol. 37, no. 2, pp. 1168–1174, 2017.
[57] X. Wang *et al.*, "Trichostatin A, a histone deacetylase inhibitor, reverses epithelial–mesenchymal transition in colorectal cancer SW480 and prostate cancer PC3 cells," *Biochem. Biophy. Res. Commun.*, vol. 456, no. 1, pp. 320–326, 2015.
[58] L. Tessitore, A. Davit, I. Sarotto, and G. Caderni, "Resveratrol depresses the growth of colorectal aberrant crypt foci by affecting bax and P21 CIP expression," *Carcinogenesis*, vol. 21, no. 8, pp. 1619–1622, 2000.
[59] J. C. Marsh *et al.*, "The influence of drug interval on the effect of methotrexate and fluorouracil in the treatment of advanced colorectal cancer," *J. Clin. Oncol.*, vol. 9, no. 3, pp. 371–380, 1991.
[60] C. B. Burness and S. T. Duggan, "Trifluridine/Tipiracil: A review in metastatic colorectal cancer," *Drugs*, vol. 76, no. 14, pp. 1393–1402, 2016.

[61] F. Amiri, A. H. Zarnani, H. Zand, F. Koohdani, M. Jeddi-Tehrani, and M. Vafa, "Synergistic anti-proliferative effect of resveratrol and etoposide on human hepatocellular and colon cancer cell lines," *Eur. J. Pharmacol.*, vol. 718, no. 1–3, pp. 34–40, 2013.
[62] L. B. Saltz *et al.*, "Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer," *N. Engl. J. Med.*, vol. 343, no. 13, pp. 905–914, 2000.
[63] L. S. Faried *et al.*, "Inhibition of the mammalian target of rapamycin (mTOR) by rapamycin increases chemosensitivity of CaSki cells to paclitaxel," *Eur. J. Cancer*, vol. 42, no. 7, pp. 934–947, 2006.
[64] C. M. Lee *et al.*, "Phosphatidylinositol 3-kinase inhibition by LY294002 radiosensitizes human cervical cancer cell lines," *Clin. Cancer Res.*, vol. 12, no. 1, pp. 250–256, 2006.
[65] Y. Zhao, W. You, J. Zheng, Y. Chi, W. Tang, and R. Du, "Valproic acid inhibits the angiogenic potential of cervical cancer cells via HIF-1α/VEGF signals," *Clin. Transl. Oncol.*, vol. 18, no. 11, pp. 1123–1130, 2016.
[66] M. Lim *et al.*, "Ligand-independent and tissue-selective androgen receptor inhibition by pyrvinium," *ACS Chem. Biol.*, vol. 9, no. 3, pp. 692–702, 2014.
[67] O. Alqawi, M. Moghaddas, and G. Singh, "Effects of geldanamycin on HIF-1 α mediated angiogenesis and invasion in prostate cancer cells," *Prostate Cancer Prostatic Dis.*, vol. 9, no. 2, pp. 126–135, 2006.
[68] H. C. Chang *et al.*, "Desipramine-induced apoptosis in human PC3 prostate cancer cells: Activation of jnk kinase and caspase-3 pathways and a protective role of [Ca2+] i elevation," *Toxicology*, vol. 250, no. 1, pp. 9–14, 2008.
[69] A. Imrali, X. Mao, M. Yeste-Velasco, J. Shamash, and Y. Lu, "Rapamycin inhibits prostate cancer cell growth through cyclin D1 and enhances the cytotoxic efficacy of cisplatin," *Amer. J. Cancer Res.*, vol. 6, no. 8, pp. 1772–1784, 2016.
[70] R. V. Roy, S. Suman, T. P. Das, J. E. Luevano, and C. Damodaran, "Withaferin A, a steroidal lactone from Withania somnifera, induces mitotic catastrophe and growth arrest in prostate cancer cells," *J. Natural Products*, vol. 76, no. 10, pp. 1909–1915, 2013.
[71] J. Gilloteaux, J. M. Jamison, D. Neal, and J. L. Summers, "Synergistic antitumor cytotoxic actions of ascorbate and menadione on human prostate (DU145) cancer cells in vitro: Nucleus and other injuries preceding cell death by autoschizis," *Ultrastruct. Pathol.*, vol. 38, no. 2, pp. 116–140, 2014.
[72] V. Singh, P. K. Jaiswal, I. Ghosh, H. K. Koul, X. Yu, and A. De Benedetti , "Targeting the TLk1/NEK1 DDR axis with thioridazine suppresses outgrowth of androgen independent prostate tumors," *Int. J. Cancer*, vol. 145, no. 4, pp. 1055–1067, 2019.
[73] Y. Meng *et al.*, "Natural BH3 mimetic (-)-gossypol chemosensitizes human prostate cancer via Bcl-xL inhibition accompanied by increase of puma and noxa," *Mol. Cancer Ther.*, vol. 7, no. 7, pp. 2192–2202, 2008.
[74] K. Kuroda *et al.*, "Activated Akt prevents antitumor activity of gefitinib in renal cancer cells," *Urology*, vol. 74, no. 1, pp. 209–215, 2009.
[75] B. R. Seo *et al.*, "Anisomycin treatment enhances TRAIL-mediated apoptosis in renal carcinoma cells through the down-regulation of Bcl-2, c-FLIP (L) and Mcl-1," *Biochimie*, vol. 95, no. 4, pp. 858–865, 2013.
[76] M. Zakikhani, R. Dowling, I. G. Fantus, N. Sonenberg, and M. Pollak, "Metformin is an AMP kinase–dependent growth inhibitor for breast cancer cells," *Cancer Res.*, vol. 66, no. 21, pp. 10 269–10 273, 2006.
[77] E. A. Mandujano-Tinoco , J. C. Gallardo-P érez, A. Mar ín-Hernández, R. Moreno-S ánchez, and S. Rodr íguez-Enríquez, "Anti-mitochondrial therapy in human breast cancer multi-cellular spheroids," *Biochimica et Biophysica Acta (BBA)-Mol. Cell Res.*, vol. 1833, no. 3, pp. 541–551, 2013.
[78] R. C. Coombes *et al.*, "Danazol treatment of advanced breast cancer," *Cancer Treatment Rep.*, vol. 64, no. 10–11, pp. 1073–1076, 1980.
[79] M. Mody *et al.*, "Rosiglitazone sensitizes MDA-MB-231 breast cancer cells to anti-tumour effects of tumour necrosis factor-α, CH11 and CYC202," *Endocrine-Related Cancer*, vol. 14, no. 2, pp. 305–315, 2007.
[80] R. K. Srivastava, R. Kurzrock, and S. Shankar, "MS-275 sensitizes TRAIL-resistant breast cancer cells, inhibits angiogenesis and metastasis, and reverses epithelial-mesenchymal transition in vivo," *Mol. Cancer Ther.*, vol. 9, pp. 3254–3266, 2010.
[81] P. G. GuerreroJr *et al.*, "Synthesis of arotinoid acid and temarotene using mixed (Z)-1, 2-bis (organylchalcogene)-1-alkene as precursor," *Tetrahedron Lett.*, vol. 53, no. 39, pp. 5302–5305, 2012.
[82] R. S. Mehta *et al.*, "Combination anastrozole and fulvestrant in metastatic breast cancer," *N. Engl. J. Med.*, vol. 367, no. 5, pp. 435–444, 2012.
[83] P. Nie, W. Hu, T. Zhang, Y. Yang, B. Hou, and Z. Zou, "Synergistic induction of erlotinib-mediated apoptosis by resveratrol in human non-small-cell lung cancer cells by down-regulating survivin and up-regulating puma," *Cellular Physiol. Biochem.*, vol. 35, no. 6, pp. 2255–2271, 2015.
[84] D. Zgouras, U. Becker, S. Loitsch, and J. Stein, "Modulation of angiogenesis-related protein synthesis by valproic acid," *Biochem. Biophys. Res. Commun.*, vol. 316, no. 3, pp. 693–697, 2004.
[85] C. Habold *et al.*, "Trichostatin a causes p53 to switch oxidative-damaged colorectal cancer cells from cell cycle arrest into apoptosis," *J. Cellular Mol. Med.*, vol. 12, no. 2, pp. 607–621, 2008.
[86] Q. Ji *et al.*, "Resveratrol inhibits invasion and metastasis of colorectal cancer cells via MALAT1 mediated Wnt/β-catenin signal pathway," *PLoS One*, vol. 8, no. 11, 2013, Art. no. e78700.
[87] A. A. Momtazi-borojeni , E. Abdollahi, F. Ghasemi, M. Caraglia, and A. Sahebkar, "The novel role of pyrvinium in cancer therapy," *J. Cellular Physiol.*, vol. 233, no. 4, pp. 2871–2881, 2018.
[88] J. Jones *et al.*, "The histone deacetylase inhibitor valproic acid alters growth properties of renal cell carcinoma in vitro and in vivo," *J. Cellular Mol. Med.*, vol. 13, no. 8b, pp. 2376–2385, 2009.
[89] J. Y. Park, P. Y. Lin, and R. H. Weiss, "Targeting the PI3k–AKT pathway in kidney cancer," *Expert Rev. Anticancer Therapy*, vol. 7, no. 6, pp. 863–870, 2007.

Fei Wang received the bachelor's degree in microelectronics from Northwestern Polytechnical University, in 2011, and the master's degree in computer science and technology from Shaanxi Normal University, in 2016. He is currently working toward the PhD degree with the Division of Biomedical Engineering at the University of Saskatchewan. His research topic is drug repositioning.

Yulian Ding received the BSc degree in computer science and technology from Luoyang Normal University, China, in 2014, and the MSc degree in computer science and technology from Shananxi Normal University, China, in 2017. Currently, she is working toward the PhD degree with the Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, Canada. Her current research interests include biomolecular network analytics, machine learning in bioinformatics, and big biological/clinical data analytics.

Xiujuan Lei (Member, IEEE) received the PhD degree from Northwestern Polytechnical University, in 2005. She has been a professor in Shaanxi Normal University, since 2013. She was a visiting scholar with the State University of New York at Buffalo, from 2009 to 2010. She was a visiting scholar in Peking University, from 2013 to 2014. Her research interests include bioinformatics, data mining, and swarm intelligence computing.

Bo Liao received the PhD degree in computational mathematics from the Dalian University of Technology, Dalian, China, in 2004. He is currently working with Hainan Normal University as a professor, and the dean of School of Mathematics and Statistics. He worked in the Graduate University of Chinese Academy of Sciences as a Postdoctorate from 2004 to 2006. His current research interests include bioinformatics, data mining, and machine learning.

Fang-Xiang Wu (Senior Member, IEEE) received the bachelor's and master's degrees in applied mathematics from the Dalian University of Technology, in 1990 and 1993, respectively the first PhD degree in control theory and its applications from Northwestern Polytechnical University, in 1998, and the second PhD degree in bioinformatics and computational biology from the University of Saskatchewan (U of S), in August 2004. During September 2004 August 2005, he worked as a PDF with the Laval University Biomedical Research Center. He is currently a full professor with three Departments (Computer Science, Biomedical Engineering, and Mechanical Engineering) at the University of Saskatchewan. His research interests include artificial intelligence, machine/deep learning, computational biology and bioinformatics, medical image analytics, and complex network analytics. He is serving as the editorial board member of five international journals and as the guest editor of numerous international journals, and as the program committee chair or member of many international conferences.