This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

A Self-Training Subspace Clustering Algorithm under Low-Rank Representation for Cancer Classification on Gene Expression Data

Chun-Qiu Xia, Ke Han, Yong Qi, Yang Zhang, and Dong-Jun Yu

Abstract—Accurate identification of the cancer types is essential to cancer diagnoses and treatments. Since cancer tissue and normal tissue have different gene expression, gene expression data can be used as an efficient feature source for cancer classification. However, accurate cancer classification directly using original gene expression profiles remains challenging due to the intrinsic high-dimension feature and the small size of the data samples. We proposed a new self-training subspace clustering algorithm under low-rank representation, called SSC-LRR, for cancer classification on gene expression data. Low-rank representation (LRR) is first applied to extract discriminative features from the high-dimensional gene expression data; self-training subspace clustering (SSC) method is then used to generate the cancer classification predictions. The SSC-LRR was tested on two separate benchmark datasets in control with four state of the art classification methods. It generated cancer classification predictions with an overall accuracy 89.7% and a general correlation 0.920, which are 18.9% and 24.4% higher than that of the best control method respectively. In addition, several genes (RNF114, HLA-DRB5, USP9Y and PTPN20) were identified by SSC-LRR as new cancer identifiers that deserve further clinical investigation. Overall, the study demonstrated a new sensitive avenue to recognize cancer classifications from large-scale gene expression data.

Index Terms—Cancer classification, gene expression data, low-rank representation, self-training, semi-supervised learning, subspace clustering

1 Introduction

Cancer is a major and serious public health problem worldwide. It is the second leading cause of death in the United States and about 1,600 Americans died from cancer per day in 2016 [1]. Since the accurate identification of cancer types plays an essential role in both cancer diagnosis and prognosis, cancer classification has become an important field of cancer research [2]. Traditional approaches to cancer classification rely on the subjective interpretation of clinical and histopathological information [3], which can lead to variable and uncertain results in the clinical diagnoses and prognoses even for the same cancer patient, because of the subjective interpretations and doctors' personal experiences [4].

With the development of high-throughput cancer detection technology, large volumes of cancer data have been quickly accumulated. Among the molecular-level cancer data, gene expression is one of the most commonly used resources for cancer classification, due to the fact that the tumor tissues often have specific pattern different from normal tissues in the gene expression [5]. Most of the quantitative cancer classification approaches on gene expression are based on machine learning. According to the specific techniques that are used, the machine-learning approaches can be grouped into three categories of unsupervised, supervised and semi-supervised methods. The unsupervised methods, which do not utilize the label information, are often used to classify unlabeled data. Many unsupervised clustering algorithms, such as K-means, the finite mixture of Gaussians and hierarchical clustering, have been successfully applied to cancer gene expression data [6]. Recently, more sophisticated unsupervised methods, including graph regularized subspace segmentation based on nonnegative matrix factorization (NMF) [7] and Gauss-Seidel based NMF [8], were proposed to improve the accuracy of the classifications.

Compared with unsupervised methods, the supervised methods tend to provide a more precise classification by training a model on a labeled dataset. Support vector machine (SVM) [9, 10] is one of the most commonly used supervised learning techniques for the classification of gene expression data. For example, Piao et al. proposed an ensemble correlation-based gene selection algorithm using SVM to perform cancer classification [11]. Liu et al. developed a cancer classification method that combines principal component analysis (PCA) and SVM training [12]. Other supervised learning models, such as total principal component regression [13], decision tree model [14] and random forest [15], have

Chun-Qiu Xia, Ke Han, Yong Qi, and Dong-Jun Yu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei, Nanjing 210094, China. (e-mail: billxia2012@yahoo.com, hanke@njust.edu.cn, qyong@njust.edu.cn, njyudj@njust.edu.cn)

Yang Zhang is with the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA. (e-mail: zhng@umich.edu)

Manuscript submitted 13 Feb. 2017.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

been also investigated for gene expression-based cancer classifications.

Recently, much attention has been paid to semi-supervised methods due to the accumulation of large volume of unlabeled gene expression data, which cannot be well utilized by traditional supervised learning methods [16]. For example, Cai et al. proposed a semi-supervised dimensionality reduction based on random subspace segmentation for cancer classification [17]. Halder and Misra presented a semi-supervised fuzzy k-Nearest Neighbor algorithm based on the self-training technique for cancer classification [18]. These studies demonstrated usefulness of the unlabeled data by the semi-supervised methods.

Although progress has been made in gene expression-based cancer classification, challenges remain to the machine learning approaches. First, gene expression data is intrinsically a high-dimensional data, which often has tens of thousands of feature components to train; Second, the size of most publicly available gene expression dataset, i.e., the number of samples, is usually small; Third, labeled gene expression data accounts for only a small fraction of total data; those large volume of unlabeled data contains valuable information for cancer classification and deserves deep exploration. All these aspects make the analysis of gene expression data a typical high-dimensional and small sample size problem with unlabeled data instances [19].

Considerable effort has been made to meet the challenges. Dimensionality reduction algorithms have been proposed to deal with the high-dimension and small size problems of the data samples. For example, nonnegative PCA was used to analyze the latent structures contained in the high-dimensional data with lower dimensional features [20]; Sharma et al. proposed a top-r feature selection algorithm to overcome the shortcoming of conventional feature selection algorithm [21]. There have also emerged many semi-supervised methods for exploiting the information buried in unlabeled data. For example, label propagation, which utilizes the unlabeled data by a label propagation mechanism based on the hypothesis that similar data should have similar labels, has been applied to gene expression-based classification [22]. Semi-supervised projective non-negative matrix factorization (Semi-PNMF) is also a semi-supervised learning method, which jointly learns a non-negative subspace from concatenated labeled and unlabeled samples, and has demonstrated its potential for cancer classification [4].

In this study, we aim to make further progress in dealing with these challenges. A new composite method, called SSC-LRR, which integrates self-training subspace clustering (SSC) [23] and low-rank representation (LRR) [24], is proposed by considering the three characteristics of gene expression data, i.e., the high-dimensionality, the small sample size, and the existence of unlabeled data. Compared to other cancer classification approaches, the main merit of SSC-LRR is in two aspects: (1) LRR is introduced to perform subspace segmentation that can reduce the dimension of the gene expression data; and (2) an enhanced semi-supervised self-training subspace clustering algorithm based on LRR can effectively utilize both the labeled and unlabeled gene expression data. The testing results on several publicly available benchmark gene expression datasets showed that a composite approach by SSC-LRR can achieve more accurate cancer classification with the accuracy outperforming the state of the art methods.

2 Material And Methods

2.1 Benchmark datasets

Two gene expression datasets were used to benchmark the methods of this study. The first is called GCM, which was created by Ramaswamy et al. [3] and is publicly available at http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi. GCM consists of the expression profiles of 218 tumor samples representing 14 common human cancer types, and each sample contains 16,030 gene expression values. The authors further divided the GCM dataset into three subsets: a training subset of 144 samples, a testing subset of 54 samples, and a subset of 20 poorly differentiated samples (tumors). Considering that the poorly differentiated samples may induce a biased evaluation result, only the well-differentiated samples, i.e., 144 samples in training subset and 54 samples in testing subset, were combined to form a benchmark dataset in this study. This reduced GCM dataset is termed as R-GCM in the subsequent description. To remove the noise of very low values and the saturation effects of very high values in the R-GCM dataset, Zhang et al. [4] performed a pre-processing step by placing the gene expression data into a specific box constraint ranging from 20 to 16,000 units and then excluding those genes whose ratios across samples are under 5 and absolute variations across samples are under 500, respectively. After the pre-processing step, 11,370 genes remain which are eventually used in this study.

The second dataset is MBD, which consists of 76 medulloblastoma samples representing 4 cancer subgroups [25], i.e., WNT-subgroup, SHH-subgroup, subgroup-3, and subgroup-4. We removed the samples that are unlabeled or outliers, and the final MBD dataset contains 73 medulloblastoma samples, each of which contains 54,675 gene expression values.

Supplementary Table S1 summarizes the detailed composition of the R-GCM and MBD datasets. We noticed that the magnitudes of expression values vary significantly between different sets of gene expression data. In light of this, we rescaled each gene expression value into the range of (0, 1) by using a sigmoid function, and the rescaled gene expression data were then used to benchmark cancer classification methods.

2.2 Self-training subspace clustering under LRR

In this study, we propose a novel semi-supervised self-training clustering algorithm under low-rank representation (SSC-LRR), which integrates both the advantages of low-rank representation and self-training. Particularly, we utilize an efficient data selection procedure to relieve the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

mistake-reinforcement problem of self-training. A detailed description of the concepts of LRR [24] is provided in the Supplementary Section 2.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbf{R}^{d \times n}$ be an original data matrix, among which each column is the d-dimensional feature vector of a sample (gene expression data in this study), and n is the total number of samples. Suppose that only part of samples is labeled, a SSC-LRR pipeline is designed as follows:

Step 1: Perform LRR on original data matrix $\mathbf{X}$. First, we apply LRR to the original data matrix $\mathbf{X}$ with Supplementary (S5), and the decomposed low-rank matrix $\mathbf{Z}$ and sparse matrix $\mathbf{E}$ are obtained.

The matrix $\mathbf{Z}$ can be divided into a labeled matrix $\mathbf{Z}_l = \left[\mathbf{z}_1^l, \mathbf{z}_2^l, \cdots, \mathbf{z}_p^l\right] \in \mathbf{R}^{n \times p}$ and an unlabeled matrix $\mathbf{Z}_u = \left[\mathbf{z}_1^u, \mathbf{z}_2^u, \cdots, \mathbf{z}_{n-p}^u\right] \in \mathbf{R}^{n \times (n-p)}$ according to the labels of samples. By re-arranging the orders of samples, $\mathbf{Z}$ can be rewritten as $\mathbf{Z} = [\mathbf{Z}_l \quad \mathbf{Z}_u]$. Similarly, matrix $\mathbf{E}$ can also be divided into $\mathbf{E}_l$ and $\mathbf{E}_u$, and be rewritten as $\mathbf{E} = [\mathbf{E}_l \quad \mathbf{E}_u]$.

Step 2: Perform K-means clustering algorithm on $\mathbf{Z}$ and $\mathbf{E}$, respectively. The key problem for performing K-means is how to initialize the central points of clusters. Taking $\mathbf{Z} = [\mathbf{Z}_l \quad \mathbf{Z}_u]$ as an example, the initial point of cluster (class) i can be determined by

$$\mathbf{p}^{(i)} = \frac{\sum_{j=1}^{n_l^{(i)}} \mathbf{z}_{l,j}^{(i)}}{n_l^{(i)}} \tag{1}$$

where $\mathbf{p}^{(i)}$ is the initial point of cluster i, $n_l^{(i)}$ is the number of points of the cluster i in $\mathbf{Z}_l$, $\mathbf{z}_{l,j}^{(i)}$ is the j-th sample in $\mathbf{Z}_l$ belonging to the cluster i with $1 \le i \le C$, and C is the total number of clusters (classes).

Based on the initial central points of clusters obtained by (1), we perform the standard K-means algorithm on matrix $\mathbf{Z}$ until each of the unlabeled samples is clustered into one of the C clusters. According to the clustering results on $\mathbf{Z}$, the labels of those unlabeled samples are predicted. The predicted labels of $\mathbf{Z}_u$, together with the labels of $\mathbf{Z}_l$, form the label vector of $\mathbf{Z}$, denoted as $\mathbf{l}_\mathbf{Z}$. This procedure can be formulated as follows:

$$\mathbf{l}_\mathbf{Z} = \text{K-means}(\mathbf{Z}, \text{dist}_\mathbf{Z}) \tag{2}$$

where $\text{dist}_\mathbf{Z}$ denotes the distance metric used for clustering $\mathbf{Z}$.

Similarly, we can obtain the label vector of $\mathbf{E}$, denoted as $\mathbf{l}_\mathbf{E}$, by using the same procedure of obtaining $\mathbf{l}_\mathbf{Z}$, i.e.,

$$\mathbf{l}_\mathbf{E} = \text{K-means}(\mathbf{E}, \text{dist}_\mathbf{E}) \tag{3}$$

where $\text{dist}_\mathbf{E}$ denotes the distance metric used for clustering $\mathbf{E}$.

The K-means algorithm outline above can be easily extended, with any other appropriate distance scales, to facilitating different application scenarios of data clustering problems.

Step 3: Select unlabeled samples as labeled ones for next round clustering. After obtaining the clustering results, i.e., $\mathbf{l}_\mathbf{Z}$ and $\mathbf{l}_\mathbf{E}$, we need to decide which unlabeled samples should be selected and used as labeled samples for the next round clustering.

An unlabeled sample, say sample i in $\mathbf{Z}_u$, will be selected as the labeled data for the next round clustering if and only if

$$l_{\mathbf{z}_i} = l_{\mathbf{e}_i} \tag{4}$$

where $l_{\mathbf{z}_i}$ is the predicted label of the unlabeled sample i (i.e., $\mathbf{z}_i^u$) according to the clustering results of $\mathbf{l}_\mathbf{Z}$, and $l_{\mathbf{e}_i}$ is the predicted label of the unlabeled sample i (i.e., $\mathbf{e}_i^u$) according to the clustering results of $\mathbf{l}_\mathbf{E}$. All the unlabeled samples satisfying (4) constitute the set of chosen samples, denoted as S, for next round clustering.

ALGORITHM 1
PIPELINE OF THE SEMI-SUPERVISED SELF-TRAINING CLUSTERING ALGORITHM UNDER LOW-RANK REPRESENTATION

Input	$\mathbf{X}$ - The original data matrix; λ - The control parameter of LRR; $\text{dist}_\mathbf{Z}$, $\text{dist}_\mathbf{E}$ - The distance metrics of K-means for clustering $\mathbf{Z}$ and $\mathbf{E}$, respectively; maxIterNum - The max number of iteration.
Output	$\mathbf{l}_\mathbf{Z}$ - The clustering results, among which the labels of unlabeled samples are predicted.
Step 1:	Perform LRR on original data matrix $\mathbf{X}$.
	$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_{2,1}$; $s.t.,\ \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}$ Re-arrange $\mathbf{Z}$ and $\mathbf{E}$ as $\mathbf{Z} = [\mathbf{Z}_l \quad \mathbf{Z}_u]$ and $\mathbf{E} = [\mathbf{E}_l \quad \mathbf{E}_u]$, respectively; currentIterNum $\leftarrow 0$; // Counter of clustering iterations
Step 2:	Perform K-means algorithm on $\mathbf{Z}$ and $\mathbf{E}$, respectively.

	$\mathbf{l}_\mathbf{Z} = \text{K-means}(\mathbf{Z}, \text{dist}_\mathbf{Z})$; //Using (1) to determine the initial point of each cluster.
	$\mathbf{l}_\mathbf{E} = \text{K-means}(\mathbf{E}, \text{dist}_\mathbf{E})$; // Using a method similar to (1) to determine the initial point of each cluster
	// $\mathbf{l}_\mathbf{Z}$ and $\mathbf{l}_\mathbf{E}$ are the clustering results on $\mathbf{Z}$ and $\mathbf{E}$, respectively;
	currentIterNum $\leftarrow$ currentIterNum $+ 1$;

Step 3:	Select unlabeled samples as labeled ones for next round clustering.

	$S \leftarrow \Phi$;
	FOR each unlabeled sample i in $\mathbf{Z}_u$
	Let $l_{\mathbf{z}_i}$ and $l_{\mathbf{e}_i}$ be the predicted labels of sample i according to the clustering results of $\mathbf{l}_\mathbf{Z}$ and $\mathbf{l}_\mathbf{E}$, respectively.
	IF $l_{\mathbf{z}_i} = l_{\mathbf{e}_i}$
	$S \leftarrow S \cup \{i\}$; //select an unlabeled sample
	END IF
	END FOR

Step 4:	Decide whether to terminate the algorithm, and update $\mathbf{Z}$ and $\mathbf{E}$.

	IF $S = \varnothing$ or currentIterNum > maxIterNum :
	RETURN $\mathbf{l}_\mathbf{Z}$;
	ELSE
	FOR each selected unlabeled sample i in S
	move $\mathbf{z}_i^u$ from $\mathbf{Z}_u$ to $\mathbf{Z}_l$;
	move $\mathbf{e}_i^u$ from $\mathbf{E}_u$ to $\mathbf{E}_l$;
	END FOR
	$\mathbf{Z} \leftarrow [\mathbf{Z}_l \quad \mathbf{Z}_u]$; //update $\mathbf{Z}$
	$\mathbf{E} \leftarrow [\mathbf{E}_l \quad \mathbf{E}_u]$; //update $\mathbf{E}$
	GOTO Step 2;
	END IF

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Step 4: Decide whether to terminate the algorithm, and update $\mathbf{Z}$ and $\mathbf{E}$. If $S = \Phi$ or current iteration number is greater than the predefined iteration number, the procedure is stopped with $\mathbf{I_Z}$ returned as the final clustering results. Otherwise, $\mathbf{Z}$ and $\mathbf{E}$ are updated as follows:

For each selected unlabeled sample i in S, we update $\mathbf{Z}_l$ and $\mathbf{Z}_u$ by moving $\mathbf{z}_i^u$ from $\mathbf{Z}_u$ to $\mathbf{Z}_l$. Similarly, we update $\mathbf{E}_l$ and $\mathbf{E}_u$ by moving $\mathbf{e}_i^u$ from $\mathbf{E}_u$ to $\mathbf{E}_l$. Further, the updated $\mathbf{Z}_l$ and $\mathbf{Z}_u$ will be merged as new $\mathbf{Z}$, and the updated $\mathbf{E}_l$ and $\mathbf{E}_u$ will be merged as new $\mathbf{E}$, for next round clustering. After this update step, the procedure goes to Step 2 for next round clustering.

The detailed procedures of the proposed SSC-LRR are summarized in Algorithm 1.

2.3 Evaluation indexes and validation protocol

The prediction results of a multi-class predictor can be represented as a $C \times C$ confusion matrix, denoted as $M = [m_{ii}]^{C \times C}$, where m_{ij} represents the number of cases in which a sample in class i was predicted as that in class j, and C is the total number of classes [26]. The Recall (*REC*), False Positive Rate (*FPR*), and Matthews Correlation Coefficient (*MCC*) are calculated by

$$\begin{aligned} REC(i) &= TP(i)/(TP(i)+FN(i)) \\ FPR(i) &= FP(i)/(TN(i)+FP(i)) \\ MCC(i) &= (TP(i)\cdot TN(i) - FP(i)\cdot FN(i))/ \\ &\sqrt{(TP(i)+FP(i))\cdot(TP(i)+FN(i))\cdot(TN(i)+FP(i))\cdot(TN(i)+FN(i))} \end{aligned} \tag{5}$$

where $TP(i)$, $FP(i)$, $TN(i)$, and $FN(i)$ represent the true positive, false positive, true negative, and false negative rates for class i, respectively.

The Overall Accuracy (Q) and Generalized Correlation (*GC*) are calculated by [26, 27]:

$$\begin{cases} Q = \sum_{i=1}^{C} TP(i)/N \\ GC = \sqrt{\dfrac{\sum_{i=1}^{C}\sum_{j=1}^{C}\dfrac{m_{ij}-e_{ij}}{e_{ij}}}{N(C-1)}} \end{cases} \tag{6}$$

where $N = \sum_{i=1}^{C}\sum_{j=1}^{C} m_{ij}$ is the total number of samples and $e_{ij} = (TP(i)+FN(i)) \times (TP(j)+FP(j))/N$ is the expected number of cases in cell (i, j) of the confusion matrix under the null hypothesis assumption that there is no correlation between assignments and predictions.

To examine the efficiency of the methods, we first randomly split each of the benchmark datasets (R-GCM and MBD) into separate training and testing subsets. The samples in the training subset are considered as labeled data, with those in the testing subset as unlabeled. Then, we applied learning models to the datasets (i.e., the union of training and testing subsets) to obtain the predicted labels of the unlabeled samples in the testing subset. The procedure is repeated independently for 100 times to eliminate randomness, with the average prediction performance reported.

Note that in this validation protocol, a crucial issue is how to split a benchmark dataset into training and testing subsets. For fair comparison, unless otherwise stated, the M is set to 6 for evaluating all the considered methods, because the numbers of samples for most cancer types are close to 12 (cf. Supplementary Table S1).

3 Results And Discussion

3.1 LRR and self-training help improve cancer classification performance

To examine the effects of LRR and self-training to gene expression-based cancer classification, we performed a comparison between K-means, K-means with PCA, K-means with LRR, and SSC-LRR (i.e., K-means with LRR and self-training) on the R-GCM dataset using the validation protocol in Section 2.3, where the parameters in the models are decided by a protocol descripted in Supplementary Section 3. For a fair comparison, we also applied the same procedure for determining the initial points of clusters to the three compared methods. Table 1 summarizes the detailed performance comparisons between the three methods with $M = 6$ for splitting benchmark dataset.

The results in Table 1 showed that the proposed SSC-LRR outperformed K-means, K-means with PCA, and K-means with LRR methods on both R-GCM and MBD datasets regarding all the five evaluation indexes. On the R-GCM dataset, for example, the overall accuracy Q and the generalized correlation *GC*, which are two overall performance indexes, of K-means are improved by 6.2% (=(0.734-0.691)/0.691) and 5.0%, respectively, by incorporating LRR into K-means (i.e., K-means with LRR). This observation demonstrates that LRR can extract more discriminative features from high-dimensional data. We also found that the Q and *GC* values can be further improved by 2.5% and 1.7%, respectively, by incorporating self-training into K-means with LRR (i.e., SSC-LRR), which demonstrates the positive effects of unlabeled data in self-training. Similar (or even more) improvements were observed on the MBD dataset, where Q and *GC* are improved by 17.2% and 10.0%, respectively, by incorporating self-training into K-means with LRR in this dataset. We found that the performance of K-means with PCA is even worse than K-means. A possible reason is that PCA abandons too much information hidden in original high dimensional feature components.

In Table 1, the values in parentheses are the p-values in the student's t-test between SSC-LRR and other methods. Most of the p-values are far below 0.05 (except for the REC in relation to K-means with LRR method on R-GCM dataset), suggesting that the improvement by SSC-LRR is statistically significant.

3.2 Comparisons with other cancer classification methods

In this section, we compare SSC-LRR with several widely used supervised and semi-supervised cancer classification methods,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 1
COMPARISONS OF DIFFERENT METHODS FOR CANCER CLASSIFICATION ON THE R-GCM AND MBD DATASETS*

Datasets	Method	*REC*	*FPR*	*MCC*	*Overall Acc* (*Q*)	*General Corr* (*GC*)
R-GCM	K-means	0.647 (1.72×10^{-5})	0.031 (1.69×10^{-6})	0.639 (3.04×10^{-6})	0.691 (2.38×10^{-6})	0.713 (2.20×10^{-6})
	K-means with PCA	0.471 (4.11×10^{-6})	0.056 (1.11×10^{-6})	0.458 (1.39×10^{-6})	0.532 (2.42×10^{-7})	0.633 (1.40×10^{-5})
	K-means with LRR	0.706 (3.76×10^{-1})	0.026 (6.49×10^{-4})	0.678 (3.63×10^{-2})	0.734 (6.23×10^{-3})	0.749 (3.13×10^{-3})
	SSC-LRR	0.715	0.023	0.762	0.753	0.762
MBD	K-means	0.846 (8.42×10^{-3})	0.112 (2.32×10^{-3})	0.717 (3.74×10^{-4})	0.644 (2.58×10^{-3})	0.833 (7.16×10^{-3})
	K-means with PCA	0.675 (1.14×10^{-7})	0.171 (6.94×10^{-6})	0.532 (3.68×10^{-7})	0.459 (7.24×10^{-9})	0.698 (3.22×10^{-6})
	K-means with LRR	0.762 (6.62×10^{-4})	0.079 (1.32×10^{-4})	0.770 (5.05×10^{-5})	0.765 (8.93×10^{-5})	0.836 (4.69×10^{-5})
	SSC-LRR	0.959	0.033	0.888	0.897	0.920

TABLE 2
COMPARISONS OF SSC-LRR WITH OTHER METHODS FOR CANCER CLASSIFICATION ON THE R-GCM AND MBD DATASETS *

Datasets	Method	*REC*	*FPR*	*MCC*	*Overall Acc* (*Q*)	*General corr* (*GC*)
R-GCM	SVM	0.510 (1.79×10^{-7})	0.052 (1.25×10^{-4})	0.503 (5.55×10^{-7})	0.554 (3.37×10^{-5})	0618 (1.34×10^{-5})
	RPCA+SVM	0.521 (1.20×10^{-4})	0.051 (3.79×10^{-6})	0.512 (3.52×10^{-6})	0.562 (8.96×10^{-7})	0.624 (1.48×10^{-6})
	Label Propagation	0.324 (5.26×10^{-6})	0.056 (2.85×10^{-4})	0.308 (2.96×10^{-7})	0.409 (2.34×10^{-4})	0.445 (2.79×10^{-4})
	Semi-PNMF	0.658 (8.07×10^{-4})	0.029 (8.64×10^{-5})	0.633 (9.03×10^{-5})	0.707 (3.17×10^{-5})	0.710 (5.10×10^{-4})
	SSC-LRR	0.715	0.023	0.762	0.753	0.762
MBD	SVM	0.636 (1.56×10^{-5})	0.167 (3.70×10^{-3})	0.504 (1.34×10^{-3})	0.471 (1.34×10^{-4})	0.651 (7.67×10^{-4})
	RPCA+SVM	0.451 (1.16×10^{-8})	0.175 (6.92×10^{-6})	0.279 (5.85×10^{-8})	0.653 (6.35×10^{-8})	0.385 (2.28×10^{-8})
	Label Propagation	0.553 (9.99×10^{-5})	0.184 (8.94×10^{-4})	0.348 (3.77×10^{-5})	0.646 (7.85×10^{-3})	0.415 (1.09×10^{-4})
	Semi-PNMF	0.884 (4.62×10^{-2})	0.106 (3.96×10^{-2})	0.635 (9.78×10^{-3})	0.708 (3.49×10^{-2})	0.676 (1.58×10^{-3})
	SSC-LRR	0.959	0.033	0.888	0.897	0.920

*Parameter M for splitting benchmark dataset is set to 6. Values in parentheses are p-values in student t-test relative to SSC-LRR.

including two supervised methods (Support Vector Machine (SVM) [9, 10] and RPCA+SVM [12]), and two semi-supervised methods (Label Propagation [22], Semi-supervised Projective Non-negative Matrix Factorization (Semi-PNMF) [4]), for cancer classification on the two benchmark datasets with the same validation protocol.

3.2.1Performance comparisons on R-GCM dataset

The upper part of Table 2 lists the performance comparisons between SVM, RPCA+SVM, Label Propagation, Semi-PNMF, and SSC-LRR, on the R-GCM dataset with the parameter $M = 6$ for splitting benchmark dataset. Several observations can be made from the data in Table 2.

First, SVM has been recognized as a powerful supervised classification method for dealing with small sample size and high-dimensional problems. However, it is found that the two considered semi-supervised methods, i.e., SSC-LRR and Semi-PNMF performed significantly better than SVM with an improvement of 19.9% and 15.3% on *Q*, and 14.4% and 9.2% on *GC*, respectively. A possible reason is that the extremely severe conflict between sample size and sample dimensionality (i.e., 198 vs. 11,370) seriously deteriorated the performance of SVM, where Semi-PNMF and SSC-LRR effectively utilized the information buried in unlabeled samples and can thus achieve a better classification performance than SVM.

Second, the proposed SSC-LRR outperformed RPCA+SVM with an improvement of 9.0% and 13.8% on *Q* and *GC*, respectively. The underlying reasons may be as follows: in RPCA+SVM, a robust PCA is used to obtain a sparse matrix from the original data and only the sparse matrix is then used to identify key genes for training the SVM model. While in SSC-LRR, LRR is applied to decompose the original data matrix into a LRR matrix and a sparse matrix. Importantly, both the LRR matrix, which encodes the intrinsic structures of gene expression data, and the sparse matrix, which encodes key genes, are utilized for training a prediction model (see Section 3.3 below). On the other hand, compared with RPCA+SVM, SSC-LRR has the capability of learning from unlabeled data by introducing a self-training procedure. These two aspects account for the major contributions to the observed improvement of SSC-LRR over RPCA+SVM.

Third, it was found that SSC-LRR outperformed Semi-PNMF, which is a most recently proposed semi-supervised model specially designed for cancer classification on gene expression data. SSC-LRR achieved the best performance on R-GCM with the highest *Q* (0.753) and *GC* (0.762), which are 4.6% and 0.52%, respectively, higher than that of the second-best performer (Semi-PNMF). This observation further demonstrates the efficacy of the SSC-LRR for performing cancer classification on gene expression data.

Finally, Label Propagation is a semi-supervised method that also utilized unlabeled samples as Semi-PNMF and SSC-LRR do. However, we found it performed even worse than SVM that only uses the labeled samples. In Label Propagation, samples are represented as points and there is an edge between every two points. The weight of the edge is calculated based on the Euclidean distance between two points. As described in Supplementary Section 3, it is found that the Euclidean distance metric is not a good choice for reflecting the distribution of gene expression data. We speculate that this is probably the reason that accounts for the poor performance of Label Propagation.

In Table 2, the *p*-values between SSC-LRR and other methods are all below 10^{-3}, suggesting that the difference between SSC-LRR and the other control methods is statistically significant.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

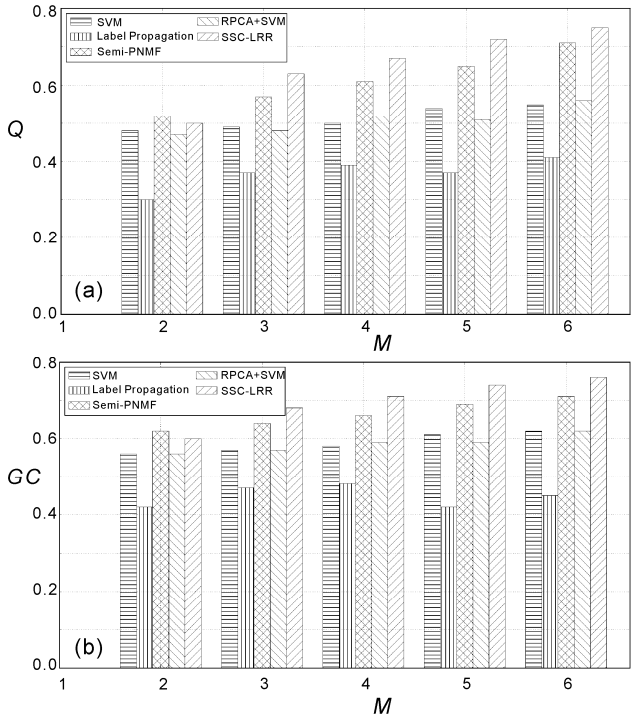


Fig. 1. Performance comparisons between SVM, RPCA+SVM, Label Propagation, Semi-PNMF, and SSC-LRR under different values of parameter M on R-GCM dataset. (a) Comparison results regarding Q; (b) comparison results regarding GC.

We also investigated the effects of the parameter *M* (i.e., how many samples of each class are selected to constitute the training subset) on the classification performances by varying the values of *M* from 2 to 6 with a step size of 1. Figs. 1 (a) and (b) list the performances regarding *Q* and *GC*, respectively, of the five methods considered under different values of parameter *M*.

The data in Fig. 1 showed a slight dependence of the performances of the methods on *M*. It is a common knowledge that the performance of a learning model increases with the number of labeled data used for training. Clearly, the trend of the performance variation is consistent with this knowledge, although there exist slight fluctuations for Label Propagation. Meanwhile, the relative performance of different methods is largely consistent, where SSC-LRR outperformed the control methods when $M > 2$, demonstrating the robustness of SSC-LRR. However, we found that the accuracy of SSC-LRR is slightly lower than that of Semi-PNMF when $M = 2$. The reason is that SSC-LRR generates mistaken reinforcement when *M* is low due to the too few initial training samples, although it has been designed to eliminate mistaken reinforcement during the iterative learning procedure.

3.2.2Performance comparisons on the MBD dataset

To comprehensively investigate the efficacy of the proposed SSC-LRR, we made a further performance comparison of SSC-LRR with the four control methods on the MBD dataset, the result of which is summarized in the lower part of Table 2.

In the MBD dataset, the conflict between sample size and sample dimensionality (73 vs. 54,675) is even more severe than that (198 vs. 11,370) of the R-GCM dataset. Nevertheless, it was found again that the proposed SSC-LRR achieved a better performance than the control methods, regarding both the three averaged binary indexes (i.e., *REC*, *FPR*, and *MCC*) and the two global indexes (i.e., *Q* and *GC*). The Overall Accuracy ($Q = 0.897$) and General Correlation ($GC = 0.920$) by SSC-LRR are 18.9% and 24.4%, respectively, higher than that of the second-best performer (Semi-PNMF).

Interestingly, the RPCA+SVM method, which used RPCA to extract key genes for classification, performed much worse than SVM itself regarding *REC*, *FPR*, *MCC*, and *GC*. We analyzed the detailed classification results of RPCA+SVM and found a large number of samples belonging to WNT, SHH, and G3 cancers have been mistakenly classified into G4 cancer, leading to a lower performance than that of SVM. This is probably due to the dominantly high sample of G4 cancers in the MBD dataset, where the number of samples of G4 cancer is about 2~5 times as much as that of other three cancers of WNT, SHH, and G3; therefore, the RPCA+SVM model trained on the MBD dataset has been biased to the majority cancer type (G4).

3.3 Why does SSC-LRR achieve better performance?

Taking the R-GCM dataset as an example, the gene expression data can be represented by a matrix of $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbf{R}^{11370\times198}$, among which each column is a sample. Without loss of generality, the samples in $\mathbf{X}$ are arranged according to their class labels in an ascending order, i.e., the label vector of $\mathbf{X}$ is $\boldsymbol{I}_x = [1,1,\cdots,1,2,2,\cdots,2,13,13,\cdots,13,14,14,\cdots,14] \in \boldsymbol{R}^{198}$, where the class labels of 1, 2, ..., and 14 represent the cancer types of BR, PR, ⋯, and CNS, respectively (cf. Table S1). Then, the data matrix $\mathbf{X}$ can be decomposed into a low-rank representation matrix $\mathbf{Z} \in \mathbf{R}^{198\times198}$ and a sparse matrix $\mathbf{E} \in \mathbf{R}^{11370\times198}$ by using (S5) in Supplementary Information. Each matrix can be displayed as an image by regarding each element in the matrix as a pixel value. Considering that there exists minus values in the decomposed matrix $\mathbf{Z}$ and $\mathbf{E}$, we linearly scaled the elements in $\mathbf{Z}$ and $\mathbf{E}$ into range of (0,1) for drawing images. In addition, we scaled the heights of $\mathbf{X}$ and $\mathbf{E}$ to the size of 198 for better visual effects. Fig. 2 illustrates the color images of the three matrices, $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{E}$, respectively.

3.3.1LRR matrix unveils the intrinsic structures for different cancer types and subtypes

The image in Fig. 2b shows that there are multiple rectangles, highlighted with yellow border along the diagonal of the image $\mathbf{Z}$, which contain relatively higher pixel values for each type of the 14 cancers (In MBD dataset, there will be 4 rectangles for 4 subtypes of medulloblastoma). These rectangles indicate that samples belonging to the same class often have the same subspace structure. In other words, the low-rank representation matrix $\mathbf{Z}$ can unveil the intrinsic structure of data much better

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

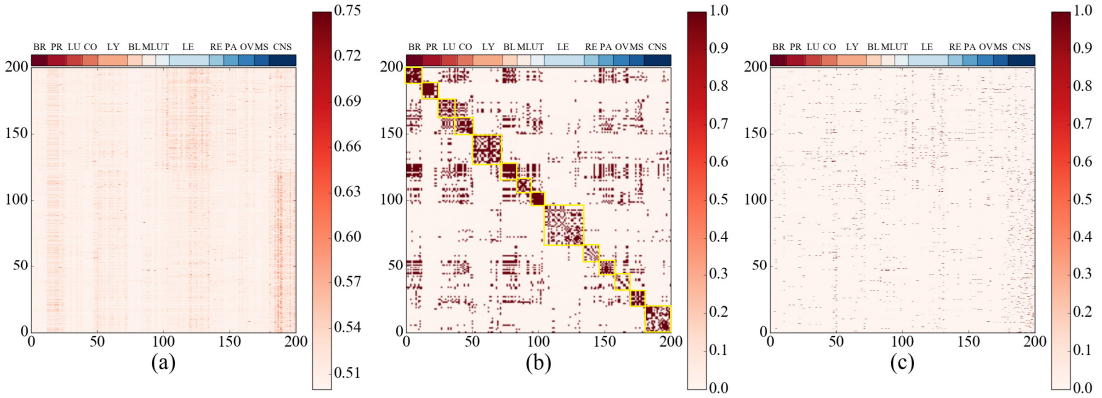


Fig. 2. An intuitive explanation of why SSC-LRR can achieve better performance on gene expression data. (a) Image of original data matrix X; (b) image of low-rank representation matrix Z; (c) image of sparse matrix E.

than the original data matrix $\mathbf{X}$. Accordingly, the low-rank representation matrix $\mathbf{Z}$ can provide more useful discriminative information than $\mathbf{X}$, leading to a better classification performance. From a biological point of view, different types of cancers are often associated with some specific genes and therefore the corresponding gene expression data may fall into specific feature subspaces, which can be unveiled by using LRR.

3.3.2Sparse matrix encodes the key genes for discriminating cancer types and subtypes

A critical step in the SSC-LRR is to utilize both the LRR matrix and the sparse matrix E for the selection of unlabeled samples as labeled ones (cf. (3) and (4)). The reason why we use the sparse matrix is that it encodes key genes for discriminating cancer types.

As displayed in Fig. 2c, there are few nonzero values in sparse matrix $\mathbf{E}$, and each nonzero value in $\mathbf{E}$ partially reflects the up- or down-regulated genes of differential expressions. Since each row of the sparse matrix $\mathbf{E}\#$ corresponds to a specific gene, we can calculate the significance of each gene by summing up the absolute values in the corresponding row of the sparse matrix $\mathbf{E}$. Generally, the higher the significance of a gene is, the more important the gene will be. According to the calculated significances of genes, we can locate the most influential genes by choosing those genes with larger significances. Table 3 lists the top 10 most influential genes and their related cancer types regarding the R-GCM dataset.

Among the top 10 identified key genes, 8 of them have been verified by previous studies to be related to one or several cancers considered in R-GCM dataset. For example, MEG3 [28] , which can activate the expression of the TP53 [29], and EXT1 [30] are tumor suppressors and their loss has been found in various types of tumors, including CNS, LE, LU. JUNB is a target gene of miRNA-149*, which can influence the cell cycle and proliferation capacity of T-ALL (T-cell acute lymphoblastic leukemia, a subtype of LE) cells [31]. MYC is overexpressed in many types of tumors, such as LE, LY, and regulates the expression of two immune checkpoint proteins on the tumor cell surface [32]. CFL1 is highly related to the invasiveness of the malignancy of the central nervous system (CNS) [33]. RABIF transcripts are overexpressed in pancreatic cancer (PA) tissues compared to normal pancreas [34]. DBI controls lung cancer (LU) progression by regulating β-oxidation [35]. MRPL48 is a candidate diagnostic marker and tumor suppressor for adrenocortical carcinoma [36].

It is interesting that two key genes, i.e., RNF114 and HLA-DRB5, which were identified from the sparse matrix E, have not yet been reported to relate with any caner types considered in R-GCM. Nevertheless, both play an important role in immune systems reported by existing studies: RNF114 regulates the immune responses and may involve in T-cell activation [37]; HLA-DRB5 plays a central role in the immune

TABLE 3
THE TOP 10 MOST INFLUENTIAL GENES AND THEIR RELATED CANCER TYPES IDENTIFIED ON R-GCM DATASET

No. of Genes*	Official Symbol†	Related Cancer Types
5128	RNF114	Potentially related
7179	MEG3	CNS, LU, etc.
3916	JUNB	LE
3648	MYC	LE, LY, etc.
4276	HLA-DRB5	Potentially related
3077	CFL1	CNS
1608	EXT1	LE, Non-melanoma Skin Cancer
6839	RABIF	PA
4232	DBI	LU
7180	MRPL48	Adrenocortical Carcinoma

*No. of Genes: Row no. of genes in the R-GCM data matrix.
†Official Symbol: Unique symbol provided by HGNC (HUGO Gene Nomenclature Committee).

TABLE 4
THE TOP 10 MOST INFLUENTIAL GENES IDENTIFIED ON MBD DATASET*

No. of Gene*	Official Symbol†	Related Cancer Types
16070	USP9Y	PR
4657	LINC01419	Hepatocellular carcinoma
24466	PTPN20	Multiple myeloma
19136	HLA-DRB4	LE
3477	HCG11	PR
20625	TBX1	LU, PR, CO
20701	PCDHA2	LY
19348	MAGEA3/6	Medulloblastoma
18692	TFPI2	Glioma
22453	HOXA10	Glioblastoma

*No. of Genes: Row no. of genes in the MBD data matrix.
†Official Symbol: Unique symbol provided by HGNC.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

system and is related to an autoimmune disease, systemic lupus erythematosus [38]. Considering that the two genes were identified as key genes for cancer classification, further experimental validation on their roles in cancer development will be important.

Similarly, the top 10 genes identified on MBD dataset are listed in Table 4. Among the top 10 genes, MAGEA3/6, TFPI2 and HOXA10 are highly linked to medulloblastoma and related cancers, because medulloblastoma and glioblastoma are two kinds of glioma. MAGEA3/6 is a direct target for miR-34a and is aberrantly expressed in many cancers including medulloblastoma [39]. TFPI2, a metastasis-suppressor [40], is highly expressed in low-grade gliomas, but there is no expression in high-grade gliomas, which suggests that repression of TFPI2 contributes to glioma progression [41]. DNA methylation at key regulatory CpGs in HOXA10 is associated with HOX-signature expression in glioblastoma [42].

The other identified genes have not been found responsible for medulloblastoma, but they are cancer-related and may play important roles in medulloblastoma progression. For example, USP9Y is a biomarkers of prostate cancer (PR) [43]. LINC01419 is a long non-coding RNA and is significantly overexpressed in hepatocellular carcinoma [44]. PTPN20 is a potential cancer-testis antigens in multiple myeloma [45]. The frequency of HLA-DRB4 increased significantly in male-patients with childhood acute lymphoblastic leukemia (LE) [46]. HCG11 is also a long non-coding RNA and its low expression level may predict a poor prognosis in prostate cancer (PR) [47]. TBX1 is overexpressed in various types of tumor compared to normal adult tissues, including LU, PR, and CO [48]. The down-regulation of PCDHA2 is responsible for hemophilic cell adhesion and its expression level is related to Burkitt lymphoma (LY) [49]. Among these genes, USP9Y and PTPN20 are male-specific, which may be responsible for that medulloblastoma is more prevalent in males [25].

In summary, the three characteristics of the proposed SSC-LRR, i.e., the intrinsic structures unveiled by low-rank representation matrix, the key genes encoded by sparse matrix, and the capability of learning from unlabeled data, made the major contributions to the performance of SSC-LRR on the gene expression-based cancer classifications.

4 Conclusions

The high-dimensionality, small sample size, and enrichment of unlabeled samples of the gene expression data represent three major obstacles for the machine-learning approaches to cancer classifications. To overcome these difficulties, we proposed a novel semi-supervised self-training subspace clustering algorithm based on low-rank representation, called SSC-LRR, for quantitative cancer classification using gene expression data. Here, LRR is introduced to relieve the conflictions between high-dimensionality and small sample size data features, by the extraction of the intrinsic structure of gene expression data, which are then encoded into low-dimensional discriminative features. An image-based analysis in Fig. 2 reveals that LRR does have the ability to identify the low-dimensional structure of gene expression data, which is of important benefit to the cancer classification task.

To utilize the information from the unlabeled gene expression data, self-training technique (cf. Supplementary Section 4) was applied in SSC-LRR. However, traditional self-training method is prone to reinforcing model mistakes, termed as mistake-reinforcement [50-52]. In light of this, a new efficient sample selection procedure (cf. Algorithm 1) has been developed to relieve the mistake-reinforcement problem of the traditional self-training methods.

The efficiency of LRR and self-training has been examined by the step-wise incorporation with the baseline K-means clustering algorithm. The experiment results demonstrated the capability of utilizing unlabeled gene expression data. The performance of the SSC-LRR were further benchmarked with several state of the art supervised and semi-supervised classification methods on two separate gene expression datasets, where SSC-LRR achieved an overall accuracy 89.7% and a general correlation 0.920, which are 18.9% and 24.4% higher than that of the best control method, respectively, on the MBD dataset. Further analysis of the SSC-LRR results also shows that it has the ability to identify key genes as possible cancer classifiers. For example, USP9Y and PTPN20 are recognized as two influential genes highly related to medulloblastoma, while RNF114 and HLA-DRB5 are shown to be potential cancer identifiers that deserve further clinical investigation.

Despite of the encouraging results of SSC-LRR, there is still considerable room for further improvement. First, more efficient methods are needed for identifying the intrinsic structure of gene expression data to extract more discriminative features for cancer classification. Second, only K-means clustering algorithm was explored in this study for implementing the SSC-LRR. More efficient clustering algorithms such as spectral clustering should have the potential to further improve the classification accuracy.

Finally, the SSC-LRR algorithm is not limited to the cancer classification, as it can be readily extended for other bioinformatics applications, such as image-based protein subcellular localization [53], image-based Alzheimer's disease classification [54], and protein interaction site prediction [55], where similar issues of high-dimensionality, small sample size, and enrich of unannotated data problems exist. The applications of SSC-LRR on these issues are under progress.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 61373062), the National Institute of Health (GM083107 and GM116960), the Natural Science Foundation of Jiangsu (No. BK20141403), the Fundamental Research Funds for the Central Universities (No. 30916011327), and the National Key Research and Development Program: Key Projects of International Scientific and Technological Innovation Cooperation between Governments (No. S2016G9070). Dong-Jun Yu is the corresponding author for this paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2016," *Ca-a Cancer Journal for Clinicians,* vol. 66, pp. 7-30, Jan-Feb 2016.

[2] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng*, et al.*, "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin," *Cell,* vol. 158, pp. 929-944, Aug 14 2014.

[3] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo*, et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, pp. 15149-15154, Dec 18 2001.

[4] X. Zhang, N. Guan, Z. Jia, X. Qiu, and Z. Luo, "Semi-Supervised Projective Non-Negative Matrix Factorization for Cancer Classification," *PLoS One,* vol. 10, p. e0138814, Sep 22 2015.

[5] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray*, et al.*, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nat Genet,* vol. 14, pp. 457-60, Dec 1996.

[6] M. C. D. Souto, I. G. Costa, D. S. D. Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC Bioinformatics,* vol. 9, pp. 1-14, 2008.

[7] X. Y. Chen and C. R. Jian, "Gene expression data clustering based on graph regularized subspace segmentation," *Neurocomputing,* vol. 143, pp. 44-50, Nov 2 2014.

[8] Q. Liao, N. Guan, and Q. Zhangg, "Gauss-Seidel based non-negative matrix factorization for gene expression clustering," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2364-2368.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[11] Y. Piao, M. Piao, K. Park, and K. H. Ryu, "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data," *Bioinformatics,* vol. 28, pp. 3306-3315, Dec 2012.

[12] J. X. Liu, Y. Xu, C. H. Zheng, H. Kong, and Z. H. Lai, "RPCA-Based Tumor Classification Using Gene Expression Data," *IEEE-ACM Transactions on Computational Biology and Bioinformatics,* vol. 12, pp. 964-970, Jul-Aug 2015.

[13] Y. Tan, L. Shi, W. Tong, and C. Wang, "Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data," *Nucleic acids research,* vol. 33, pp. 56-65, 2005.

[14] K. H. Chen, K. J. Wang, M. L. Tsai, K. M. Wang, A. M. Adrian, W. C. Cheng*, et al.*, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC Bioinformatics,* vol. 15, p. 49, Feb 20 2014.

[15] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics,* vol. 7, p. 1, 2006.

[16] Y. Liu, "Active learning with support vector machine applied to gene expression data for cancer classification," *Journal of chemical information and computer sciences,* vol. 44, pp. 1936-1941, 2004.

[17] X. F. Cai, J. Wei, G. H. Wen, and Z. W. Yu, "Local and Global Preserving Semisupervised Dimensionality Reduction Based on Random Subspace for Cancer Classification," *IEEE Journal of Biomedical and Health Informatics,* vol. 18, pp. 500-507, Mar 2014.

[18] A. Halder and S. Misra, "Semi-supervised fuzzy K-NN for cancer classification from microarray gene expression data," *2014 First International Conference on Automation, Control, Energy & Systems (ACES-14),* pp. 266-270, 2014.

[19] A. F. Pour and L. A. Dalton, "Optimal Bayesian Feature Selection on High Dimensional Gene Expression Data," *2014 IEEE Global Conference on Signal and Information Processing (Globalsip),* pp. 1402-1405, 2014.

[20] X. Han, "Nonnegative principal component analysis for cancer molecular pattern discovery," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 7, pp. 537-549, Jul-Sep 2010.

[21] A. Sharma, S. Imoto, and S. Miyano, "A Top-r Feature Selection Algorithm for Microarray Gene Expression Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 9, pp. 754-764, May-Jun 2012.

[22] X. Zhu and Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation," 2003.

[23] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]," *IEEE Transactions on Neural Networks,* vol. 20, pp. 542-542, 2009.

[24] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 663-670.

[25] G. Robinson, M. Parker, T. A. Kranenburg, C. Lu, X. Chen, L. Ding*, et al.*, "Novel mutations target distinct subgroups of medulloblastoma," *Nature,* vol. 488, pp. 43-48, Aug 2 2012.

[26] D. J. Yu, X. W. Wu, H. B. Shen, J. Yang, Z. M. Tang, Y. Qi*, et al.*, "Enhancing Membrane Protein Subcellular Localization Prediction by Parallel Fusion of Multi-View Features," *IEEE Transactions on Nanobioscience,* vol. 11, pp. 375-385, Dec 2012.

[27] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics,* vol. 16, pp. 412-424, May 2000.

[28] Y. Zhou, X. Zhang, and A. Klibanski, "MEG3 noncoding RNA: a tumor suppressor," *Journal of molecular endocrinology,* vol. 48, pp. R45-R53, 2012.

[29] Q. Liao, C. Liu, X. Yuan, S. Kang, R. Miao, H. Xiao*, et al.*, "Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network," *Nucleic acids research,* vol. 39, pp. 3864-3878, 2011.

[30] S. Ropero, F. Setien, J. Espada, M. F. Fraga, M. Herranz, J. Asp*, et al.*, "Epigenetic loss of the familial tumor-suppressor gene exostosin-1 (EXT1) disrupts heparan sulfate synthesis in cancer cells," *Human molecular genetics,* vol. 13, pp. 2753-2765, 2004.

[31] S.-j. Fan, H.-b. Li, G. Cui, X.-l. Kong, L.-l. Sun, Y.-q. Zhao*, et al.*, "miRNA-149* promotes cell proliferation and suppresses apoptosis by mediating JunB in T-cell acute lymphoblastic leukemia," *Leukemia research,* vol. 41, pp. 62-70, 2016.

[32] S. C. Casey, L. Tong, Y. Li, R. Do, S. Walz, K. N. Fitzgerald*, et al.*, "MYC regulates the antitumor immune response through CD47 and PD-L1," *Science,* vol. 352, pp. 227-231, 2016.

[33] A. G. Piña-Medina, V. Hansberg-Pastor, A. González-Arenas, M. Cerbón, and I. Camacho-Arroyo, "Progesterone promotes cell migration, invasion and cofilin activation in human astrocytoma cells," *Steroids,* vol. 105, pp. 19-25, 2016.

[34] F. Müller-Pillasch, F. Zimmerhackl, U. Lacher, N. Schultz, H. Hameister, G. Varga*, et al.*, "Cloning of Novel Transcripts of the Human Guanine-Nucleotide-Exchange Factor Mss4: In SituChromosomal Mapping and Expression in Pancreatic Cancer," *Genomics,* vol. 46, pp. 389-396, 1997.

[35] F. T. Harris, S. J. Rahman, M. Hassanein, J. Qian, M. D. Hoeksema, H. Chen*, et al.*, "Acyl-Coenzyme A–Binding Protein Regulates Beta-Oxidation Required for Growth and Survival of Non–Small Cell Lung Cancer," *Cancer Prevention Research,* vol. 7, pp. 748-757, 2014.

[36] G. G. Fernandez-Ranvier, J. Weng, R.-F. Yeh, D. Shibru, E. Khafnashar, K.-W. Chung*, et al.*, "Candidate diagnostic markers and tumor suppressor genes for adrenocortical carcinoma by expression profile of genes on chromosome 11q13," *World journal of surgery,* vol. 32, pp. 873-881, 2008.

[37] P. Yang, Y. Lu, M. Li, K. Zhang, C. Li, H. Chen*, et al.*, "Identification of RNF114 as a novel positive regulatory protein for T cell activation," *Immunobiology,* vol. 219, pp. 432-439, 2014.

[38] L. Wu, S. Guo, D. Yang, Y. Ma, H. Ji, Y. Chen*, et al.*, "Copy number variations of HLA-DRB5 is associated with systemic lupus erythematosus risk in Chinese Han population," *Acta biochimica et biophysica Sinica,* vol. 46, pp. 155-160, 2014.

[39] S. D. Weeraratne, V. Amani, A. Neiss, N. Teider, D. K. Scott, S. L. Pomeroy*, et al.*, "miR-34a confers chemosensitivity through modulation of MAGE-A and p53 in medulloblastoma," *Neuro-Oncology,* vol. 13, pp. 165-175, 2011.

[40] H. A. Cruickshanks, N. Vafadar-Isfahani, D. S. Dunican, A. Lee, D. Sproul, J. N. Lund*, et al.*, "Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer," *Nucleic acids research,* vol. 41, pp. 6857-6869, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2712607, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[41] S. D. Konduri, K. S. Srivenugopal, N. Yanamandra, D. H. Dinh, W. C. Olivero, M. Gujrati, *et al.*, "Promoter methylation and silencing of the tissue factor pathway inhibitor-2 (TFPI-2), in human glioma cells," *Oncogene,* vol. 22, pp. 4509-16, 2003.

[42] S. Kurscheid, P. Bady, D. Sciuscio, I. Samarzija, T. Shay, I. Vassallo, *et al.*, "Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma," *Genome Biology,* vol. 16, pp. 1-15, 2015.

[43] Y. Zhu, S. Ren, T. Jing, X. Cai, Y. Liu, F. Wang, *et al.*, "Clinical utility of a novel urine-based gene fusion TTTY15-USP9Y in predicting prostate biopsy outcome," *Urologic Oncology,* vol. 33, pp. 384.e9-384.e20, 2015.

[44] H. Zhang, C. Zhu, Y. Zhao, M. Li, L. Wu, X. Yang, *et al.*, "Long non-coding RNA expression profiles of hepatitis C virus-related dysplasia and hepatocellular carcinoma," *Oncotarget,* vol. 6, pp. 43770-43778, 2015.

[45] M. Condomines, D. Hose, T. Rème, G. Requirand, M. Hundemer, M. Schoenhals, *et al.*, "Gene expression profiling and real-time PCR analyses identify novel potential cancer-testis antigens in multiple myeloma," *Journal of Immunology,* vol. 183, pp. 832-40, 2009.

[46] M. T. Dorak, F. S. Oguz, N. Yalman, A. S. Diler, S. Kalayoglu, S. Anak, *et al.*, "A male-specific increase in the HLA-DRB4 (DR53) frequency in high-risk and relapsed childhood ALL," *Leukemia Research,* vol. 26, pp. 651-656, 2002.

[47] Y. Zhang, "Downregulation of long non-coding RNA HCG11 predicts a poor prognosis in prostate cancer," *Biomedicine & Pharmacotherapy,* vol. 83, pp. 936-941, 2016.

[48] R. I. Fernando, M. Litzinger, P. Trono, D. H. Hamilton, J. Schlom, and C. Palena, "The T-box transcription factor Brachyury promotes epithelial-mesenchymal transition in human tumor cells," *Journal of Clinical Investigation,* vol. 120, pp. 533-544, 2010.

[49] F. G. De, E. Leucci, D. Lenze, P. P. Piccaluga, P. P. Claudio, A. Onnis, *et al.*, "Gene-expression analysis identifies novel RBL2/p130 target genes in endemic Burkitt lymphoma cell lines and primary tumors," *Blood,* vol. 110, pp. 1301-7, 2007.

[50] K. Prokopiou, E. Kavallieratou, and E. Stamatatos, "An Image Processing Self-Training System for Ruling Line Removal Algorithms," *2013 18th International Conference on Digital Signal Processing (DSP),* pp. 1-6, 2013.

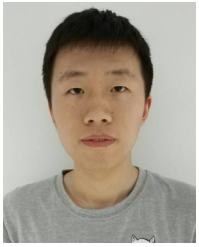
[51] X. R. Zhao, N. Evans, and J. L. Dugelay, "Semi-Supervised Face Recognition with Lda Self-Training," *2011 18th IEEE International Conference on Image Processing,* pp. 3041-3044, 2011.

[52] X. Zhu, "Semi-Supervised Learning Literature Survey," *Computer Science,* vol. 37, pp. 63-77, 2008.

[53] Y. Y. Xu, F. Yang, Y. Zhang, and H. B. Shen, "An image-based multi-label human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues," *Bioinformatics,* vol. 29, pp. 2032-40, 2013.

[54] X. Zhu, H. I. Suk, L. Wang, S. W. Lee, and D. Shen, "A novel relational regularization feature selection method for joint regression and classification in AD diagnosis," *Medical Image Analysis,* vol. 75, pp. 570-577, 2015.

[55] Z. S. Wei, K. Han, J. Y. Yang, H. B. Shen, and D. J. Yu, "Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests," *Neurocomputing,* vol. 193, pp. 201-212, 2016.



Chun-Qiu Xia received the BS degree in computer science and technology from Nanjing University, China, in 2015. Currently, he is working towards the MS degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, data mining, and pattern recognition.

Ke Han received her BS and MS degrees from Nanjing University of Science and Technology, China, in 2007 and 2009, respectively. She is currently working towards her PhD degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research interests include bioinformatics, data mining, and pattern recognition.

Yong Qi received the BS degree from East China Institute of Technology, China, in 1992, and the MS and PhD degrees from Nanjing University of Science and Technology, China, in 1999 and 2005, respectively. He is currently a professor in the School of Computer Science and Engineering, and School of Economics and Management, Nanjing University of Science and Technology, China. His research interests include machine learning, data mining, and social computing.

Yang Zhang is a professor in Department of Computational Medicine and Bioinformatics and Department of Biological Chemistry at the University of Michigan. The research interest of Dr. Zhang's lab is in protein design, protein folding, and protein structure prediction. The I-TASSER algorithm developed in his lab was ranked as one of the best methods for automated protein structure prediction in the past decade of the worldwide CASP competitions. He is the recipient of the US National Science Foundation (NSF) Career Award, the Alfred P Sloan Award, and the Dean's Basic Science Research Award, and was selected as the Thomson Reuters Highly Cited Researcher in 2015 and 2016.

Dong-Jun Yu received the PhD degree from Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2003. In 2008, he acted as an academic visitor at Department of Computer of the University of York in the UK. He also visited the Department of Computational Medicine of the University of Michigan (Ann Arbor) in 2016. He is currently a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include pattern recognition, machine learning and bioinformatics. He is the author of more than 50 scientific papers in pattern recognition and bioinformatics. He is a senior member of China Computer Federation (CCF) and a senior member of China Association of Artificial Intelligence (CAAI).