

PROCEEDINGS

Open Access

# Molecular cancer classification using a meta-sample-based regularized robust coding method

Shu-Lin Wang<sup>1,3\*</sup>, Liuchao Sun<sup>1</sup>, Jianwen Fang<sup>2,3\*</sup>

From 2013 International Conference on Intelligent Computing (ICIC 2013)  
Nanning, China. 28-31 July 2013

## Abstract

**Motivation:** Previous studies have demonstrated that machine learning based molecular cancer classification using gene expression profiling (GEP) data is promising for the clinic diagnosis and treatment of cancer. Novel classification methods with high efficiency and prediction accuracy are still needed to deal with high dimensionality and small sample size of typical GEP data. Recently the sparse representation (SR) method has been successfully applied to the cancer classification. Nevertheless, its efficiency needs to be improved when analyzing large-scale GEP data.

**Results:** In this paper we present the meta-sample-based regularized robust coding classification (MRRCC), a novel effective cancer classification technique that combines the idea of meta-sample-based cluster method with regularized robust coding (RRC) method. It assumes that the coding residual and the coding coefficient are respectively independent and identically distributed. Similar to meta-sample-based SR classification (MSRC), MRRCC extracts a set of meta-samples from the training samples, and then encodes a testing sample as the sparse linear combination of these meta-samples. The representation fidelity is measured by the  $l_2$ -norm or  $l_1$ -norm of the coding residual.

**Conclusions:** Extensive experiments on publicly available GEP datasets demonstrate that the proposed method is more efficient while its prediction accuracy is equivalent to existing MSRC-based methods and better than other state-of-the-art dimension reduction based methods.

## Introduction

With the advance of DNA microarray and next-generation sequencing (NGS) technology [1], a large amount of gene expression profiles (GEP) data has been rapidly accumulated, which requires novel analysis method to deeply mine these big data to interpret such data to gain insight into the mechanism of tumor development. Since Golub *et al.* made use of gene expression profiling data, obtained using the DNA microarray technology, to classify acute myeloid leukemia (AML) and acute lymphocytic leukemia (ALL) [2], a great number of GEP-based cancer classification methods have been proposed for classifying cancer

types or subtypes [3-6]. It has increasingly become clear that common machine learning methods such as support vector machine (SVM) [7,8] and artificial neural networks (ANN) [5,9] may not perform very well because of the curse of dimensionality, as the number of features (genes) is usually much higher than the number of samples in most GEP experiments. Therefore, the key task of GEP-based cancer classification should be the design of dimension reduction method to dramatically decrease the number of features in GEP data before constructing classification models.

Dimension reduction methods can be grouped into two categories: feature selection and feature reduction approaches. Feature selection methods [10], such as the heuristic breadth-first search algorithm, find as many optimal gene subsets as possible and further rank these genes to discover important cancer-related genes [11].

\* Correspondence: smartforestring@gmail.com; jianwen.fang@nih.gov

<sup>1</sup>College of Computer Science and Electronics Engineering, Hunan University, Hunan, 410082, China

<sup>2</sup>Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville, MD 20850, USA

Full list of author information is available at the end of the article

Feature extraction methods instead employ independent component analysis to model the gene expression data [12,13]. Gene selection methods do not alter the original representation of each gene, while feature extraction methods, which are based on projection, yield new variables that may reflect the intrinsic characteristics of original features. Other feature extraction methods such as principal component analysis (PCA)[14], linear discriminant analysis (LDA) [15], locally linear discriminant embedding (LLDE) [16], and partial least squares (PLS) [17] are also extensively applied to the dimensionality reduction of GEP. These methods can generally achieve satisfactory classification performance with the minimum dimension reduction. Both feature selection and feature extraction methods have their own advantages and disadvantages. For gene selection methods, their main advantage is that the selected genes may be related to the underlying mechanisms of cancer development. However, different gene selection methods may result in significantly different selected genes and therefore the interpretation of the results can be difficult. For the feature extraction methods, small dimension can be obtained by integrating original features. However, it is difficult to precisely interpret the biomedical meanings of derived features.

Machine learning based methods are also often called model-based methods because a predictive model is built for predicting the label of test sample. The model selection is a complex training procedure, which easily leads to over-fitting and decreased prediction performance. Recently, sparse representation (SR), a powerful data processing method that does not require model selection, has been extensively applied to face recognition [18,19] and further extended to cancer classification recently [20-22]. For example, Hang, et al. proposed a SR-based classification (SRC) method using  $l_1$ -norm minimization to classify cancer test sample. The approach models a classification problem as to find a sparse representations of test samples with respect to training samples [22]. They applied the proposed method to six cancer gene expression datasets and their experimental results demonstrated that the performance of the proposed method was comparable to or better than those of SVMs. Especially, the proposed method does not involve model selection and is robust to noise, outliers and even incomplete measurements. Zheng, et al. further presented a new SR-based method for GEP-based cancer classification, termed meta-sample-based SR classification (MSRC), where a set of meta-samples are extracted from training samples, and then a testing sample is represented as the linear combination of these meta-samples by  $l_1$ -regularized least square method [20]. Their experiments on publicly available GEP datasets have shown that MSRC is efficient for cancer classification and can achieve higher accuracy than many existing representative schemes such as SVM, SRC and least absolute

shrinkage and selection operator (LASSO) algorithm. In addition, Gan et al. proposed a new classifier, meta-sample-based robust sparse representation classifier (MRSRC) based on the MSRC method, for cancer classification [21]. Their experiments show that these methods are efficient and robust.

Previous SR-based model assumes that the coding residual follows Gaussian or Laplacian distribution, which may not be effective for describing the coding residual in practical GEP datasets, and another problem is that the sparsity constraint on coding coefficients leads to the high computational cost of SRC method. To deal with the problem, Yang et al. proposed a new coding model, namely regularized robust coding (RRC) for face recognition [23]. Here, we present a meta-sample-based regularized robust coding classification (MRRCC) method, a novel and effective cancer classification technique combining the ideas of meta-sample-based and RRC methods. A meta-sample can be represented as a linear combination of a set of training samples, which may capture the intrinsic structures of these data. The coefficient vector of a meta-sample may have only a few nonzero elements. The expression patterns over the meta-samples can reflect the gene expression patterns. Test samples belonging to the same subclass will have similar sparse representation, while different subclass would result in different sparse representations [22]. Our extensive experiments on cancer datasets show that MRRCC can achieve higher classification accuracy but with lower time complexity, compared with other SR-based methods and dimension reduction-based methods.

## Methods

### Description of SR-based problem

Let  $G = \{g_1, \dots, g_n\}$  be a set of genes and  $S = \{s_1, \dots, s_m\}$  be a set of samples.  $|G| = n$  denotes the number of genes, and  $|S| = m$  denotes the number of samples. The corresponding GEP data can be represented as a matrix  $X = (x_{ij})_{nm}$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq m$ , where  $x_{ij}$  is the expression level of gene  $g_i$  in sample  $s_j$ . Usually  $n$  is much bigger than  $m$  for a typical GEP dataset. Each vector  $s_i$  in the gene expression matrix can be regarded as a point in  $n$ -dimensional space. Each of the  $m$  columns consists of an  $n$ -element expression vector for a single sample. Let  $L = \{c_1, \dots, c_k\}$  denote the label set and  $|L| = k$  denote the number of subclasses. Because the subclass of each sample is known,  $S \times L = \{(s_i, l_i) | s_i \in R^n, l_i \in R^n, l_i \in L, i = 1, 2, \dots, m\}$  denotes the labeled sample space. The whole sample set  $X$  is randomly split into two disjoint parts: training set  $A$  and test set  $B$ . Generally, the SR-based problem could be represented as

$$\min_{\alpha} \|\alpha\|_1, \text{ s.t. } \|\gamma - A\alpha\|_2^2 \leq \varepsilon \quad (1)$$

where  $\gamma$  is a given test sample,  $A$  represents all training samples,  $\alpha$  is the coding vector of  $\gamma$  with respect to  $A$ , and  $\varepsilon$  is a small positive constant. By coding the test sample  $\gamma$  as a sparse linear combination of the training samples via Eq. (1), SR-based classifier assigns the label to the test sample  $\gamma$  based on the predictions which subclass can produce the least reconstruction error.

### Analysis flowchart of cancer GEP data

The analysis flowchart of the meta-sample-based SR method is different from those of traditional model-based and template-based methods (Figure 1). The classification models of model-based methods use the training set to predict the labels of test samples, while template-based methods create a template for each subclass using training set and then compare a test sample to the templates in order to determine the label of the test sample [3]. Although there is similarity between the analysis flowcharts of meta-sample-based SR method and template-based one, there is a major difference (Figure 1). The reconstructed test samples of the meta-sample-based SR method are relevant to not only the training set but also the original test sample, while the templates of template-based methods are only relevant to the training set. The flowchart of analysis of the meta-sample-based SR method includes five steps:

1) The whole sample set is randomly split into two disjoint parts: training set and test set, and then the meta-samples are extracted only from the training set using singular value composition (SVD).

2) The weight of each gene is calculated according to a weight function, and the genes with lower weight are removed in a test sample  $T_o$  and all meta-samples.

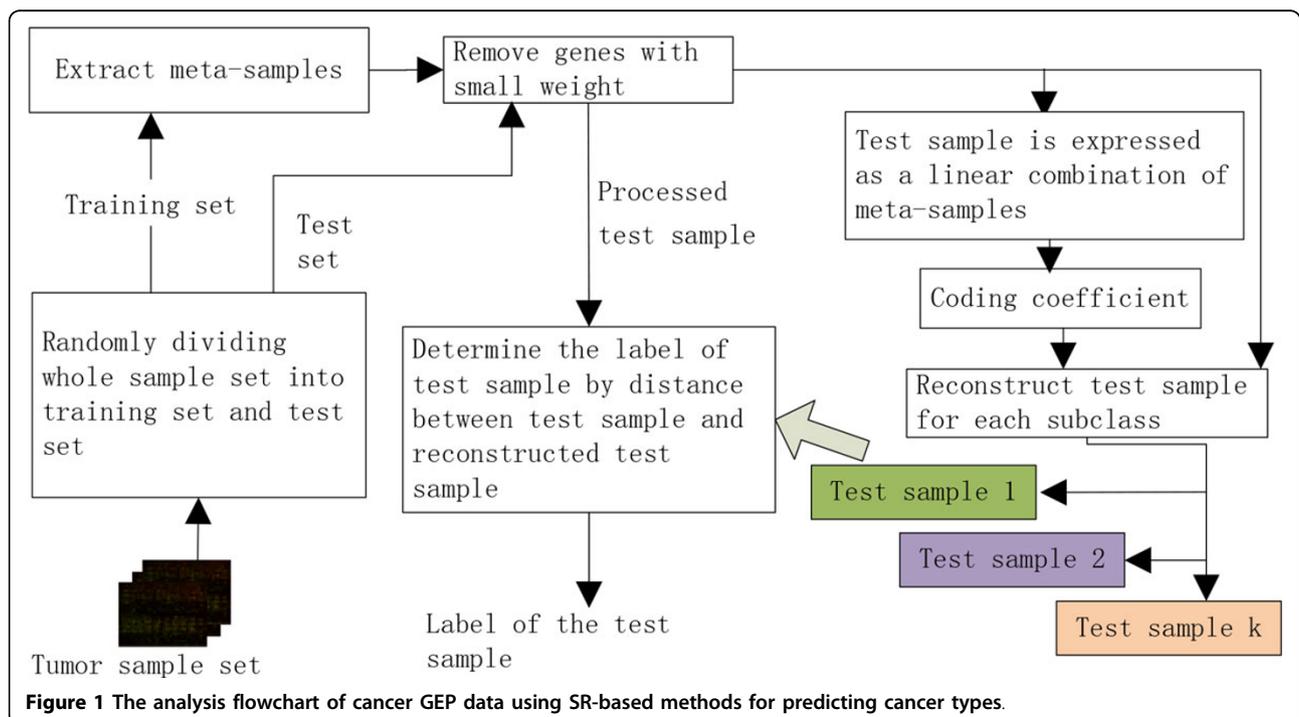
3) The test sample  $T_o$  is represented as a linear combination of all meta-samples, and the coding coefficient of the test sample  $T_o$  can be obtained by using RRC.

4) We can reconstruct the test sample for each subclass by using the meta-samples and the coding coefficient of the original test sample  $T_o$ , and the reconstructed test samples (the test sample 1, test sample 2,..., test sample  $k$ ) are denoted by  $T_1, T_2, \dots, T_k$ , where  $k$  denotes the number of subclasses in original dataset.

5) The distance between the processed test sample and each reconstructed test sample  $T_i, 1 \leq i \leq k$  is calculated, and the original test sample  $T_o$  is assigned to the subclass with minimal distance.

### Construct meta-samples

The meta-sample extracted from GEP data is commonly defined as a linear combination of all training samples. In this paper, a set of meta-sample is extracted from all training samples of one cancer type. We find that meta-sample can capture the structures to the data and offer biological insight. on the other hand, the linear combination of the meta-samples can roughly estimate genetic expression pattern of gene data [24]. Alter, *et al.* used singular value decomposition (SVD) to transform GEP data from a space of genes  $\times$  arrays to the diagonal "eigengenes  $\times$  eigenarrays" space [25], where the eigenarrays (or eigengenes) are the unique orthonormal



superpositions of the arrays (or genes). In our approach, we can decompose the gene expression data set matrix  $A$  into two matrices:

$$A_i = M_i \times V_i, \quad 1 \leq i \leq k$$

where matrix  $A_i$  is of size  $n \times m_i$ , the matrix  $M_i$  is of size  $n \times q_i$  and the matrix  $V_i$  is of size  $q_i \times m_i$ . Each of  $q_i$  columns in matrix  $M_i$  is defined as a meta-sample of the  $i$ -th subclass. Each of  $m_i$  columns in matrix  $V_i$  represents the meta-sample expression pattern of the corresponding samples.  $D = [M_1, \dots, M_i, \dots, M_k]$  denotes the constructed meta-sample set.

### Calculating coding coefficient using RRC

The meta-sample set  $D = [M_1, \dots, M_i, \dots, M_k]$  can be rewritten as  $D = [r_1; \dots; r_i; \dots; r_n]$ , where  $r_i$  denotes the expression level of the  $i$ -th gene in all meta-samples. Let  $y = [y_1; \dots; y_i; \dots; y_n]$  denote a test sample, where  $y_i$  is the expression level of the  $i$ -th gene. We can consider the cancer classification from the view point of Bayesian estimation, especially the maximum a posterior (MAP) estimation. By using the Bayesian formulation, we can calculate the coding coefficient by the following formula [23]

$$\hat{\alpha} = \operatorname{argmin}_{\alpha} \left\{ \sum_{i=1}^n \rho_{\theta}(y_i - r_i \alpha) + \sum_{j=1}^m \rho_o(\alpha_j) \right\} \quad (2)$$

where  $\rho_{\theta}(e) = -\ln f_{\theta}(e)$  and  $\rho_o(\alpha) = -\ln f_o(\alpha)$ . The coding residual  $e = y - D\alpha = [e_1; e_2; \dots; e_n]$  are with the probability density function (PDF)  $f_{\theta}(e_i)$  and the coding vector  $\alpha = [\alpha_1; \alpha_2; \dots; \alpha_m]$  are with PDF  $f_o(\alpha_j)$ . Generally, we assume that the unknown PDF  $f_{\theta}(e)$  are symmetric, differentiable and monotonic. Therefore,  $\rho_{\theta}(e)$  has following properties: (1)  $\rho_{\theta}(0)$  is the global minimal of  $\rho_{\theta}(z)$ ; (2)  $\rho_{\theta}(z) = \rho_{\theta}(-z)$ ; (3) if  $|z_1| < |z_2|$ , we can get  $\rho_{\theta}(z_1) < \rho_{\theta}(z_2)$ . Without loss of generality, we can let  $\rho_{\theta}(0) = 0$ .

There are two key issues in solving the RRC model. The first one is how to determine the distribution  $\rho_{\theta}$ . The second one is how to minimize the energy function. The RRC model in Eq. (2) can be approximated as follows.

$$\hat{\alpha} = \operatorname{argmin}_{\alpha} \left\{ \frac{1}{2} \|W^{1/2} (y - D\alpha)\|_2^2 + \sum_{j=1}^m \rho_o(\alpha_j) \right\} \quad (3)$$

where  $W$  is a diagonal matrix.

$$W_{i,i} = \omega_{\theta}(e_{0,i}) = \rho'_{\theta}(e_{0,i})/e_{0,i} \quad (4)$$

where  $W_{i,i}$  is the weight value of each gene. Thus the minimization problem of the RRC model can be transformed into calculating the diagonal weight matrix  $W$ .

The logistic function has the same properties as the hinge loss function in SVM [26], so we choose it as the weight function.

$$\omega_{\theta}(e_i) = \exp(\mu\delta - \mu e_i^2) / (1 + \exp(\mu\delta - \mu e_i^2)) \quad (5)$$

where  $\mu$  and  $\delta$  are two positive constants. Parameter  $\mu$  controls the decreasing rate from 1 to 0, and  $\delta$  controls the location of demarcation point. To make  $\omega_{\theta}(0)$  close to 1, let the value of  $\mu\delta$  be big enough. According to Eq. (4), Eq. (5) and  $\rho_{\theta}(0) = 0$ , we can get

$$\rho_{\theta}(e_i) = \frac{-1}{2\mu} (\ln(1 + \exp(\mu\delta - \mu e_i^2)) - \ln(1 + \exp(\mu\delta))) \quad (6)$$

For cancer classification, the coding coefficients associated with the dictionary atoms from the same subclass would have big absolute values. However, we do not know which subclass the testing sample will belong to. Therefore, we actually assume that the coding coefficient  $\alpha_j$  follows generalized Gaussian distribution (GGD). So we can obtain the following formula.

$$f_o(\alpha_j) = \beta \exp \left\{ -(|\alpha_j|/\sigma_{\alpha})^{\beta} \right\} / (2\sigma_{\alpha} \Gamma(1/\beta)) \quad (7)$$

where  $\Gamma$  is the gamma function.

The RRC model has two vital cases when  $\beta$  is set as two specific values [23].

When  $\beta = 1$ , GGD degenerates to Laplacian distribution, and the RRC model will become

$$\hat{\alpha} = \operatorname{argmin}_{\alpha} \left\{ \|W^{1/2} (y - D\alpha)\|_2^2 + \lambda \|\alpha\|_1 \right\} \quad (8)$$

When  $\beta = 2$ , GGD degenerates to Gaussian distribution, and the RRC model will become

$$\hat{\alpha} = \operatorname{argmin}_{\alpha} \left\{ \|W^{1/2} (y - D\alpha)\|_2^2 + \lambda \|\alpha\|_2^2 \right\} \quad (9)$$

### Iteratively reweighted regularized robust coding algorithm

Iteratively reweighted regularized robust coding (IR<sup>3</sup>C) algorithm was designed by Yang, *et al.* to solve the RRC model efficiently [23]. The overall procedure of the algorithm is as follows.

**Input:** Normalized test sample  $y$  with unit  $l_2$ -norm; meta-sample set  $D$  extracted from original training samples;  $\alpha^{(1)}$ .

**Output:**  $\alpha$

$t = 1$ ; //  $t$  denotes the iterative times.

1. Compute the gene residual  $e^{(t)} = y - D\alpha^{(t)}$

where  $\alpha^{(1)} = \left[ \frac{1}{m}; \frac{1}{m}; \dots; \frac{1}{m} \right]$ , and  $D\alpha^{(1)}$  is the mean of all meta-samples.

2. Estimate weight value of each gene as

$$\omega_{\theta} \left( e_i^{(t)} \right) = 1 / 1 + \exp \left( \mu \left( e_i^{(t)} \right)^2 - \mu \delta \right)$$

where  $\mu$  and  $\delta$  would be estimated in each iteration and  $\delta$  is associated with residual.

3. Weighted regularized robust coding coefficient:

$$\alpha^* = \operatorname{argmin}_{\alpha} \left\{ \frac{1}{2} \| (W^{(t)})^{0.5} (y - D\alpha) \|_2^2 + \sum_{j=1}^m \rho_{\sigma}(\alpha_j) \right\};$$

//Assume the  $\alpha_j$  follows generalized Gaussian distribution.

4. Update the robust coding coefficients.

If  $t = 1$ ,  $\alpha^{(t)} = \alpha^*$ ;  
 If  $t > 1$ ,  $\alpha^{(t)} = \alpha^{(t-1)} + \nu^{(t)}(\alpha^* - \alpha^{(t-1)})$ ; //where  $0 < \nu^{(t)} \leq 1$  is a suitable step size.  $\nu^{(t)}$  could be searched from 1 to 0 by the standard line-search process [27].

5. Reconstruct the test sample by coding coefficient and all meta-samples

$$y_{rec}^{(t)} = D\alpha^{(t)}, \text{ and let } t = t + 1.$$

6. Return to the step 1 until the condition of convergence  $\|W^{(t)} - W^{(t-1)}\|_2 / \|W^{(t-1)}\|_2 < \varphi$  ( $\varphi$  is a small positive scalar) is met, or reached the maximal number of iteration.

Algorithm end.

When the algorithm converges, we can use the same classification method as SRC to classify test sample.

$$\operatorname{identity}(y) = \operatorname{argmin}_d \{l_d\} \quad (10)$$

where  $l_d = \|W_{final}^{-1} (y - D_d \hat{\alpha}_d)\|_2$ ,  $D_d$  is the meta-sample set associated with  $d$ -th subclass,  $\hat{\alpha}_d$  is the final coding vector associated with  $d$ -th subclass, and  $W_{final}$  is the final weight matrix.

When  $\beta = 1$ , the time complexity of  $IR^3C$  is  $O(tm^2n)$ , where  $n$  is the number of genes,  $m$  is the number of meta-samples, and  $t$  is the iteration times. When  $\beta = 2$ , the time complexity of  $IR^3C$  is  $O(tk_1mn)$ , where  $k_1$  is the iteration number in conjugate gradient solution. The time complexity of  $IR^3C$  with  $\beta = 1$  or  $\beta = 2$  is much lower complexity than SRC whose time complexity is  $O(m^2n^{1.5})$  [23].

In literature [23] the RRC model with  $\beta = 1$  is called as RRC\_L1 and the RRC model with  $\beta = 2$  is called as RRC\_L2. However, in our method the input  $D$  of  $IR^3C$

is actually a set of meta-samples which are extracted by SVD from the original training set, so we call our methods as MRRCC1 (the meta-sample-based regularized robust coding classification 1) and MRRCC2 (the meta-sample-based regularized robust coding classification 2) corresponding to the two cases RRC\_L1 and RRC\_L2, respectively.

## Experiments

### Cancer datasets

GEP data can be obtained by two technologies, DNA microarray and next-generation sequencing (NGS) technologies. In our experiments five microarray and four NGS cancer datasets are used to evaluate the proposed method (Table 1). The five microarray datasets include Diffuse Large B-cell Lymphomas (DLBCL) [28], Acute Lymphoblastic Leukemia (ALL) [29], GCM [30], Lung cancer (Lung) [31], and MLL [32]. The DLBCL dataset contains two subclasses, i.e., DLBCL and Follicular Lymphoma (FL). The ALL dataset totally contains 248 samples that belong to six cancer subtypes: BCR-ABL, E2A-PBX1, Hyperdip>50, MLL, T-ALL and TEL-AML1. The GCM dataset consists of fourteen different cancer types. The Lung cancer dataset contains four lung cancer types and one normal tissue type (i.e., five subclasses in total). The MLL dataset contains 72 samples from three subtypes or subclasses, i.e., MLL, AML and ALL.

The four NGS datasets are downloaded from the web site: The Cancer Genome Atlas (TCGA) (<http://cancer-genome.nih.gov/>). They include Breast invasive carcinoma (called as BRCAcancer), Kidney renal clear cell carcinoma (KIRCCancer), Lung adenocarcinoma (LUADcancer), and Thyroid carcinoma (THCAcancer). All samples are matched cancer and normal tissue samples.

### Parameter selection

There are two parameters, namely  $\mu$  and  $\delta$ , in weight function Eq. (5). We calculate the  $\delta$  value as follows. Let  $l = \lfloor \tau m \rfloor$ , where  $\tau \in (0, 1)$ . Parameter  $\delta$  can be obtained by the following formula.

**Table 1 The summary of the eight cancer datasets.**

| Types      | Datasets   | #Samples | #Genes | #Subclasses(K) |
|------------|------------|----------|--------|----------------|
| Microarray | DLBCL      | 77       | 7,129  | 2              |
|            | ALL        | 248      | 12,626 | 6              |
|            | GCM        | 190      | 16,063 | 14             |
|            | Lung       | 203      | 12,601 | 5              |
|            | MLL        | 72       | 7,129  | 3              |
| NGS        | BRCAcancer | 216      | 20531  | 2              |
|            | KIRCCancer | 130      | 20531  | 2              |
|            | LUADcancer | 110      | 20531  | 2              |
|            | THCAcancer | 112      | 20531  | 2              |

$$\delta = \gamma_1(e)_l \tag{11}$$

where the vector  $e \in R^n$ ,  $\gamma_1(e)_q$  is the  $q$ -th largest element of the set  $\{e_j^2, j = 1, \dots, n\}$ . Parameter  $\mu$  is used to control the decreasing rate of the weight  $W_{i,i}$ . We can simply set  $\mu = s/\delta$ , where  $s = 8$  is defined as a constant. So the  $\delta$  value, estimated by  $\tau$  according to Eq. (11), is a very important parameter to distinguish outlier genes. The selection of parameter  $\tau$  will be further determined by our experiments.

Figure 2 shows the 10-fold cross validation prediction accuracy varying with  $\tau$  value increasing from 0.1 to 0.9 by 0.1 on four cancer datasets (two microarray datasets and two NGS datasets), from which we can see that the optimal prediction accuracy can be achieved on the four datasets when  $\tau$  is set to 0.9. So it is appropriate that  $\tau$  is fixed to 0.9 in our experiments. Furthermore, we find that the prediction accuracy on two microarray datasets is greatly affected by  $\tau$  values, while the prediction accuracy on two NGS datasets is less affected by  $\tau$  values, suggesting that the two microarray datasets are noisier than the two NGS datasets.

### Comparison with other SR-based methods

The prediction accuracy of the meta-sample-based methods is greatly affected by the number of meta-samples extracted using SVD. Figure 3 shows the prediction accuracy of the four SR-based methods (MSRC, MRSRC, MRRCC1 and MRCC2) varying with the number of meta-samples on four datasets, respectively. And

Figure 3 shows that no fixed number of meta-samples can consistently achieve the optimal performance of meta-sample-based methods. Therefore, the meta-sample-based methods require the process of optimizing the number of meta-samples. Here two-layer 10-fold cross-validation is used to evaluate the performance of the SR-based methods. The inner layer 10-fold cross-validation is used to determine the optimal number of meta-samples for training in outer layer 10-fold cross-validation, and the outer layer 10-fold cross-validation is used to evaluate the classification performance of SR-based methods. The classification accuracy obtained by five SR-based methods on the nine cancer datasets are shown in Table 2. It is clear that our methods MRRCC1 and MRRCC2 are equivalent to other three SR-based methods in optimal prediction accuracy on eight datasets except on GCM dataset.

### Comparison with dimension reduction-based methods

A two-stage method can be used to reduce the dimensionality of dataset before classification. The first stage is a process of adopting a gene filter method such as KWRST (Kruskal-Wallis rank sum test) [33] or Relief-F [34] to initially select a small set of differentially expressed genes. The second stage is a process of adopting a feature extraction method to further reduce the dimensionality of the dataset. Our previous studies have shown that the prediction accuracy of two-stage method is influenced by many factors such as normalization

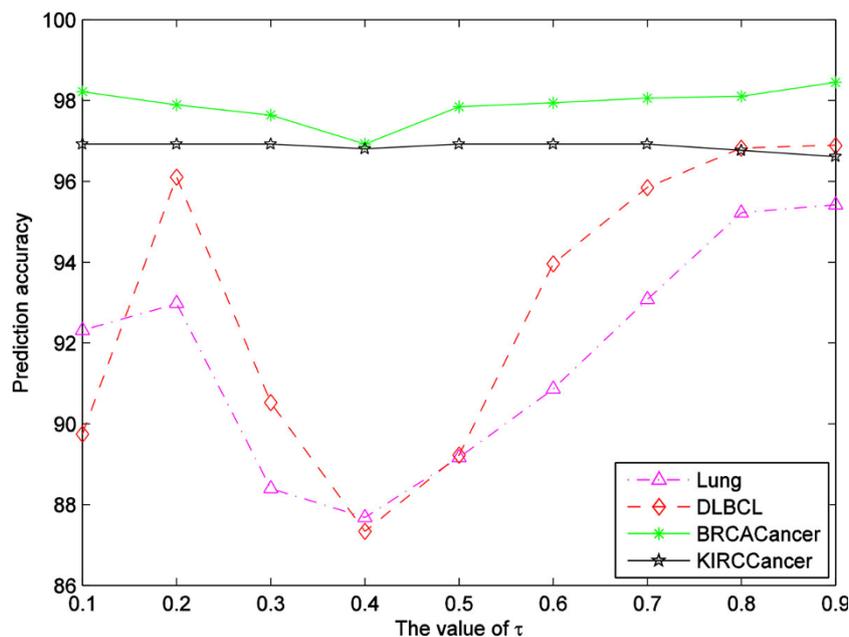


Figure 2 The prediction accuracy on the four data sets varying with different  $\tau$  value.

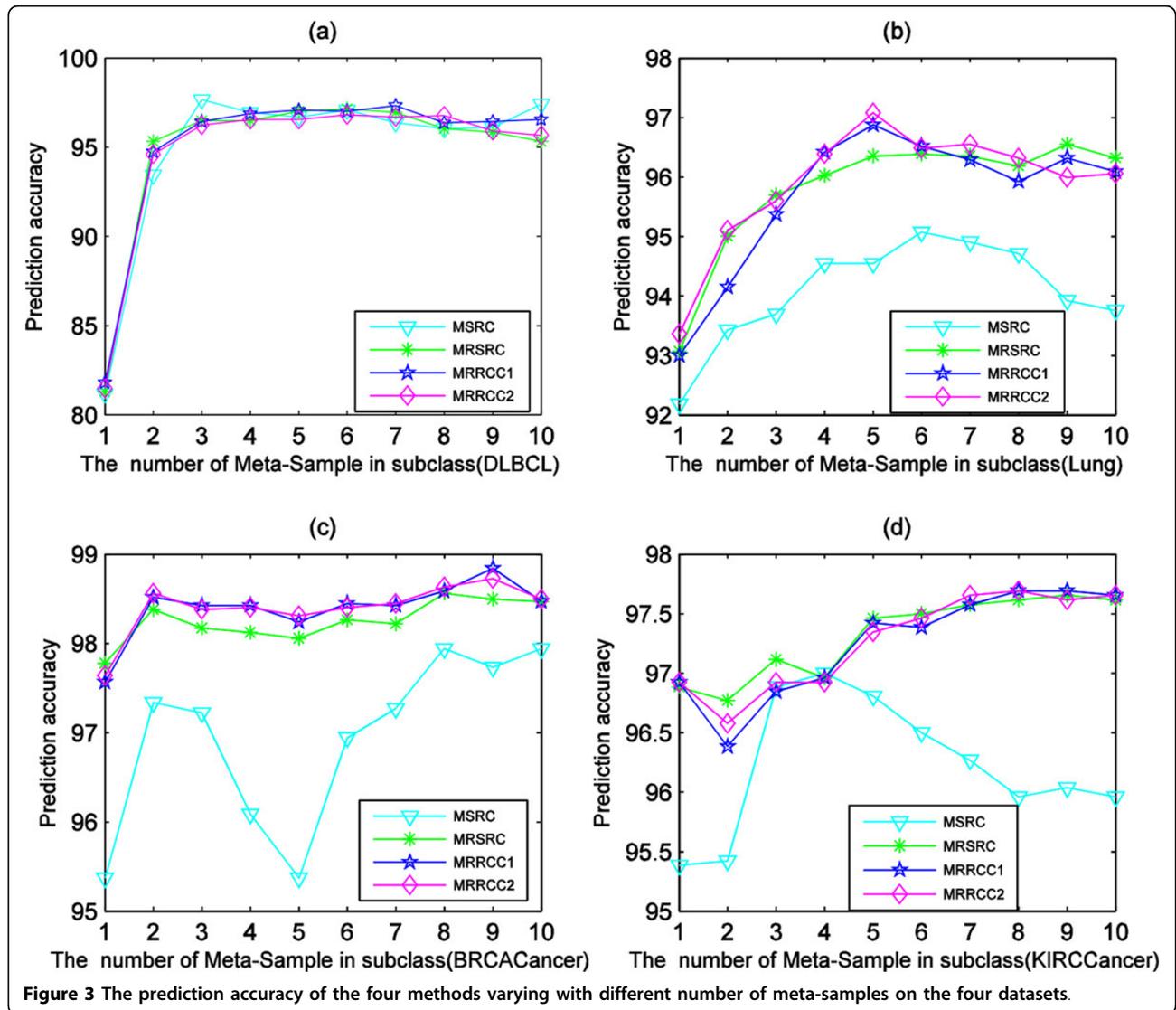


Figure 3 The prediction accuracy of the four methods varying with different number of meta-samples on the four datasets.

method, gene filter method, feature extraction method, classification method, the number of genes selected and the number of features extracted as well as different

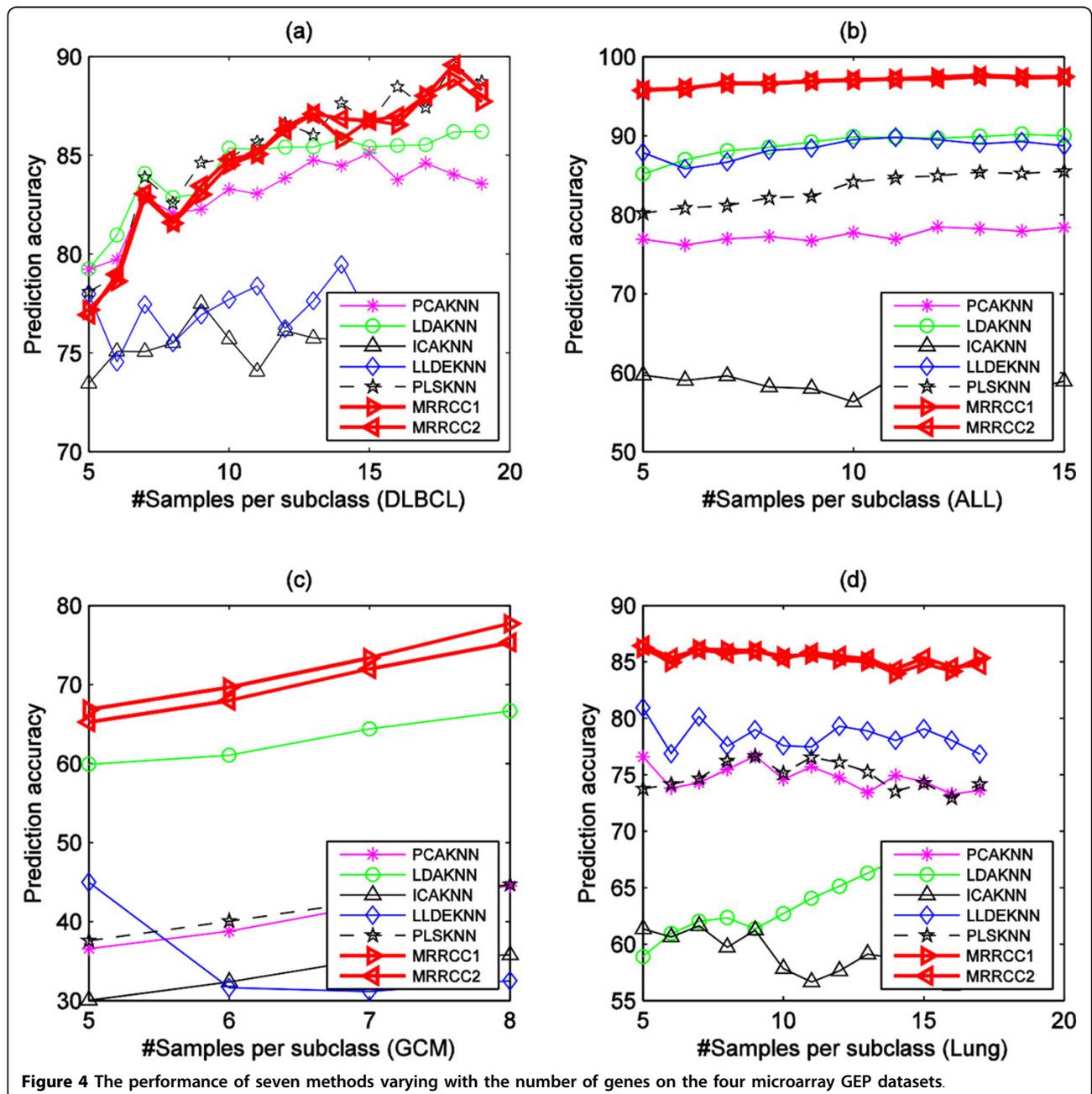
**Table 2 The classification accuracy obtained by five SR-based methods on the nine cancer datasets.**

| Types      | Datasets   | SRC   | MSRC  | MRSRC | MRRCC1       | MRRCC2       |
|------------|------------|-------|-------|-------|--------------|--------------|
| Microarray | DLBCL      | 94.75 | 96.10 | 94.81 | <b>97.40</b> | 94.81        |
|            | All        | 97.70 | 97.18 | 97.81 | 96.77        | <b>97.18</b> |
|            | GCM        | 82.93 | 82.32 | 78.79 | 79.80        | 78.79        |
|            | Lung       | 94.53 | 95.57 | 96.55 | <b>96.55</b> | <b>96.55</b> |
|            | MLL        | 96.31 | 95.83 | 98.61 | 97.22        | <b>98.61</b> |
| NGS        | BRCACancer | 96.76 | 95.83 | 99.07 | <b>99.07</b> | <b>99.07</b> |
|            | KIRCCancer | 95.92 | 95.38 | 96.92 | <b>96.92</b> | <b>96.92</b> |
|            | LUADCancer | 94.91 | 99.09 | 99.09 | <b>100</b>   | <b>99.09</b> |
|            | THCACancer | 93.30 | 87.50 | 95.54 | 92.86        | <b>95.54</b> |

division of training set and test set, etc. [35]. In our experiments, training sets and test sets are normalized by samples using the z-score normalization method. KWRST is used to filter genes and 300 top-ranked genes are initially selected. The five feature extraction methods (PCA, LDA, ICA, LLDE, and PLS) are used to reduce the dimensionality of dataset. K-nearest neighbor (KNN), one of simplest classification methods, with correlation distance is used to classify cancer samples (here 5 nearest neighbors are used). For LDA method and the datasets with two subclasses, Euclidean distance is used because only one feature is extracted. To avoid over-fitting, before classification we extract only 5 features using these feature extraction methods except LDA whose number extracted is  $K - 1$ . We call these methods as PCAKNN, LDKNN, ICAKNN, LLDEKNN, and PLSKNN, respectively.

Experiments indicate that the different divisions of training sets and test sets can also greatly affect the classification performance. In our experiments, the Balance Division Method (BDM) is used to divide each original dataset into balanced training sets and test sets [4]. For the BDM,  $Q$  samples from each subclass of the original dataset are randomly selected and used as a training set, while the remaining samples are used as test set. Here the limits of  $Q$  value ranges from 5 to  $|c_{\min}|$ , where  $c_{\min}$  denotes the subclass set with minimum number of samples in the original dataset, i.e.,  $c_{\min} = \operatorname{argmin}_{c_i}(|c_i|)$ ,  $1 \leq i \leq K$ , where

$Q$  denotes the number of subclass in dataset. We set  $Q$  value to 20 when  $|c_{\min}| > 20$ . For each  $Q$  value, the statistical mean of prediction accuracies obtained on 100 randomizations of training set and test set are calculated for each method. Figure 4 and Figure 5 show the performance of eight methods varying with different numbers of training samples per subclasses on four microarray datasets and four NGS datasets, respectively. The experimental results indicate that the performance of MRRCC1 and MRRCC2 are almost the same for all but the GCM dataset. Generally, our methods are superior to other five



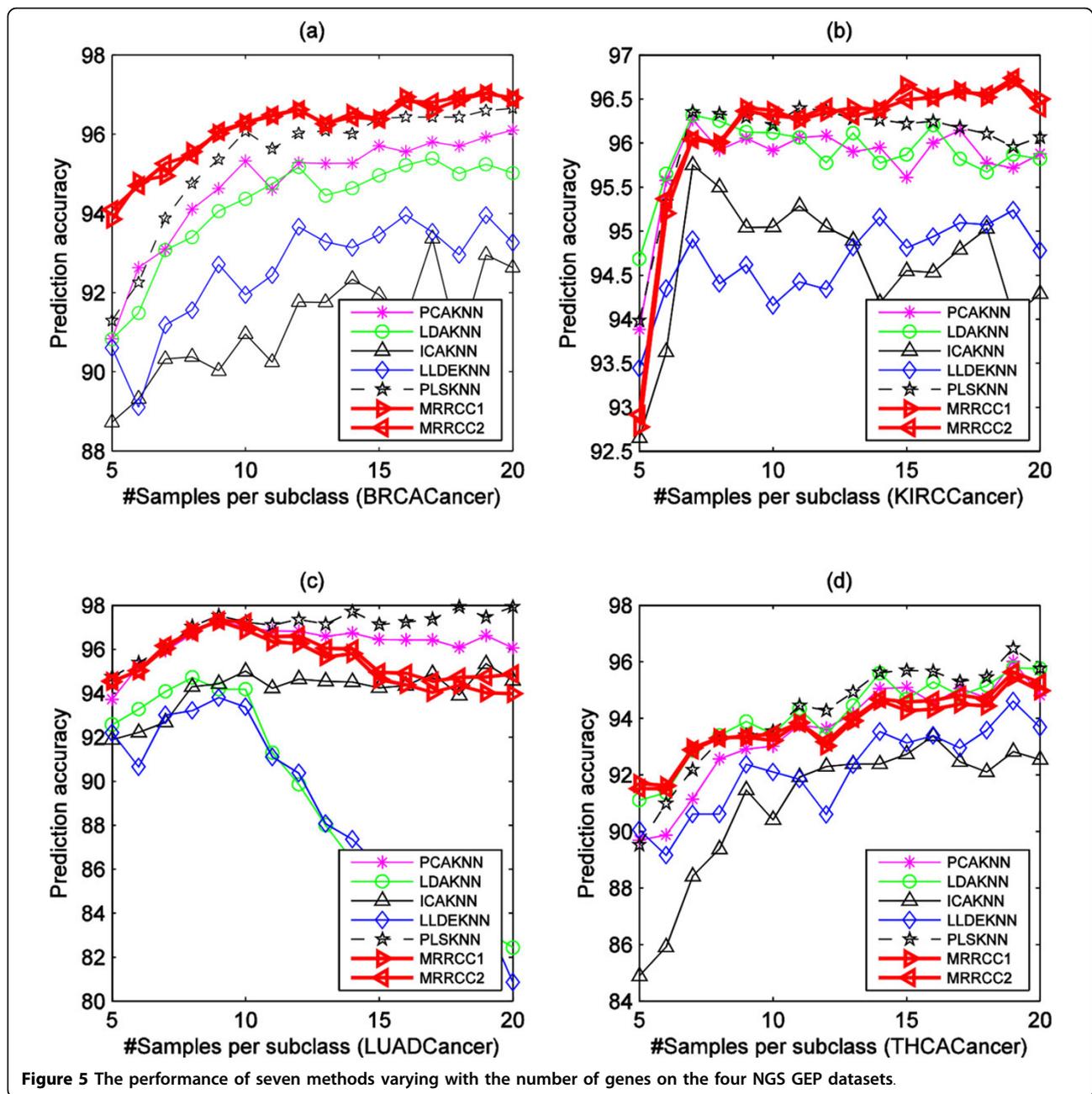


Figure 5 The performance of seven methods varying with the number of genes on the four NGS GEP datasets.

methods in prediction accuracy not only on the four microarray datasets but also on the four NGS datasets. On the LUADcancer and THCAcancer datasets the performance of our methods is slightly worse than PLSkNN in prediction accuracy when the number of the samples per subclass in training sets is greater than 10.

### Conclusions

With the development of microarray and NGS technologies, a huge amount of GEP data is rapidly accumulated, demanding more efficient analysis methods to analyze

these data. In this paper we present a novel meta-sample-based regularized robust coding for cancer classification (MRRCC) that firstly represents each test sample as a linear combination of all meta-samples which are extracted from the original training set using SVD. The coefficient vector is then obtained by  $l_2$ -regularized least square that is as powerful as  $l_1$ -norm regularization but the former has much lower computational cost [23]. The experimental results have demonstrated that MRRCC can achieve higher classification accuracy with lower computational cost than previous state-of-the-art

solutions such as SRC, MSRC and MRSRC, as well as many dimension reduction based classification methods.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

Shu-Lin Wang designed the framework of analysis, performed the partial experiments, and drafted the manuscript. Liuchao Sun also performed the partial experiments, and Jianwen Fang analyzed the numerical results and revised the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

This article was funded by the National Science Foundation of China on finding tumor-related driver pathway with comprehensive analysis method based on next-generation sequencing data and the dimension reduction of gene expression data based on heuristic method (grant nos. 61474267, 60973153 and 61133010) and the National Institutes of Health (NIH) Grant P01 AG12993 (PI: E. Michaelis).

This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 15, 2014: Proceedings of the 2013 International Conference on Intelligent Computing (ICIC 2013). The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S15>.

#### Authors' details

<sup>1</sup>College of Computer Science and Electronics Engineering, Hunan University, Hunan, 410082, China. <sup>2</sup>Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville, MD 20850, USA. <sup>3</sup>Applied Bioinformatics Laboratory, University of Kansas, Lawrence, KS 66045, USA.

Published: 3 December 2014

#### References

- Desai AN, Jere A: Next-generation sequencing: ready for the clinics? *Clin Genet* 2012, **81**(6):503-510.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 1999, **286**(5439):531-537.
- Wang SL, Fang YP, Fang JW: Diagnostic prediction of complex diseases using phase-only correlation based on virtual sample template. *Bmc Bioinformatics* 2013, **14**.
- Wang SL, Zhu YH, Jia W, Huang DS: Robust Classification Method of Tumor Subtype by Using Correlation Filters. *IEEE-Acm Transactions on Computational Biology and Bioinformatics* 2012, **9**(2):580-591.
- Wang SL, Li XL, Zhang SW, Gui J, Huang DS: Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Computers in Biology and Medicine* 2010, **40**(2):179-189.
- Zheng CH, Huang DS, Zhang L, Kong XZ: Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Transactions on Information Technology in Biomedicine* 2009, **13**(4):599-607.
- Guyon I, Weston J, Vapnik V: Gene selection for cancer classification using support vector machine. *Machine Learning* 2002, **46**(1-3):389-422.
- Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000, **16**(10):906-914.
- Xu Y, Selaru FM, Yin J, Zou TT, Shustova V, Mori Y, Sato F, Liu TC, Olaru A, Wang S, et al: Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Research* 2002, **62**(12):3493-3497.
- Saeyns Y, Inza I, Larranaga P: A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007, **23**(19):2507-2517.
- Wang SL, Li XL, Fang JW: Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. *Bmc Bioinformatics* 2012, **13**.
- Huang DS, Zheng CH: Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 2006, **22**(15):1855-1862.
- Zheng CH, Chen Y, Li XX, Li YX, Zhu YP: Tumor classification based on independent component analysis. *International Journal of Pattern Recognition and Artificial Intelligence* 2006, **20**(2):297-310.
- Wang SL, Wang J, Chen HW, Zhang BY: SVM-based tumor classification with gene expression data. *Advanced Data Mining and Applications, Proceedings* 2006, **4093**:864-870.
- Sharma A, Paliwal KK: Cancer classification by gradient LDA technique using microarray gene expression data. *Data Knowl Eng* 2008, **66**(2):338-347.
- Li B, Zheng CH, Huang DS, Zhang L, Han K: Gene expression data classification using locally linear discriminant embedding. *Computers in Biology and Medicine* 2010, **40**(10):802-810.
- Nguyen DV, Rocke DM: Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002, **18**(1):39-50.
- Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y: Robust Face Recognition via Sparse Representation. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 2009, **31**(2):210-227.
- Ma P, Yang D, Ge YX, Zhang XH, Qu Y, Huang S, Lu JW: Robust face recognition via gradient-based sparse representation. *J Electron Imaging* 2013, **22**(1).
- Zheng CH, Zhang L, Ng TY, Shiu SC, Huang DS: Metasample-based sparse representation for tumor classification. *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8**(5):1273-1282.
- Gan B, Zheng CH, Liu JX: Metasample-based robust sparse representation for tumor classification. *Engineering* 2013, **5**:78-83.
- Hang XY, Wu FX: Sparse Representation for Classification of Tumors Using Gene Expression Data. *J Biomed Biotechnol* 2009.
- Yang M, Zhang L, Yang J, Zhang D: Regularized Robust Coding for Face Recognition. *Ieee T Image Process* 2013, **22**(5):1753-1766.
- Liebermeister W: Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 2002, **18**(1):51-60.
- Alter O, Brown PO, Botstein D: Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(18):10101-10106.
- Ramsay J: The elements of statistical learning: Data mining, inference, and prediction. *Psychometrika* 2003, **68**(4):611-612.
- Hiriart-Urruty JB, Lemaréchal C: Convex analysis and minimization algorithms. Berlin; New York: Springer-Verlag; 2 1996.
- Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, et al: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 2002, **8**(1):68-74.
- Yeoh EJ RM, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002, **1**(2):133-143.
- Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al: Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(26):15149-15154.
- Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001, **98**(24):13790-13795.
- Armstrong SA, Staunton JE, Silverman LB, Pieters R, de Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics* 2002, **30**(1):41-47.
- Kruskal WH, Wallis WA: Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952, **47**(260):583-621.
- Kononenko I: Estimating attributes: Analysis and extensions of Relief. *European Conference on Machine Learning Springer-Verlag, Catania, Italy* 1994, 171-182.

35. Wang SL, You HZ, Lei YK, Li XL: **Performance Comparison of Tumor Classification Based on Linear and Non-linear Dimensionality Reduction Methods.** *Advanced Intelligent Computing Theories and Applications* 2010, **6**215:291-300.

doi:10.1186/1471-2105-15-S15-S2

**Cite this article as:** Wang *et al.*: Molecular cancer classification using a meta-sample-based regularized robust coding method. *BMC Bioinformatics* 2014 **15**(Suppl 15):S2.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

