Journal of Bioinformatics and Computational Biology
Vol. 15, No. 1 (2017) 1650037 (21 pages)

DOI: 10.1142/S0219720016500372

World Scientific
www.worldscientific.com

The method for breast cancer grade prediction and pathway analysis based on improved multiple kernel learning

Tianci Song[*,†,§], Yan Wang[†,¶], Wei Du[*,†,‖,§§], Sha Cao[†,**], Yuan Tian[*,††] and Yanchun Liang[*,‡,‡‡,§§]

[*]*College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge, Engineering of Ministry of Education Jilin University, Changchun 130012, P. R. China*

[†]*Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics University of Georgia, Athens, GA 30602, USA*

[‡]*Zhuhai Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education Zhuhai College of Jilin University, Zhuhai 519041, P. R. China*

[§]*songtc14@mails.jlu.edu.cn*
[¶]*wy6868@jlu.edu.cn*
[‖]*weidu@jlu.edu.cn*
[**]*robincaosha@gmail.com*
[††]*yuant2012@163.com*
[‡‡]*ycliang@jlu.edu.cn*

Received 29 April 2016
Revised 8 September 2016
Accepted 13 October 2016
Published 30 November 2016

Breast cancer histologic grade represents the morphological assessment of the tumor's malignancy and aggressiveness, which is vital in clinically planning treatment and estimating prognosis for patients. Therefore, the prediction of breast cancer grade can markedly elevate the detection of early breast cancer and efficiently guide its treatment. With the advent of high-throughput profiling technology, a large number of data of different types are rapidly generated, and each data provides its unique biological insight. Although many researches focused on cancer grade prediction, hardly most of them attempted to integrate multiple data types, by which we cannot only improve and boost results obtained from learning method, but also have a good understanding or explanation of biological issues. In this paper, we take advantage of a sophisticated supervised learning method called multiple kernel learning (MKL) to design a breast cancer grading predictor fusing heterogeneous data for classification of breast cancer histopathology. Furthermore, we modify our model by involving biological pathway information. The new model can evaluate the significance of various pathways in which differential expression genes fall between different breast cancer grades. The merits of the novel model are lucubration in bridging between omics data and various phenotypes of breast cancer grades, and providing an auxiliary method integrating omics data of cancer mechanism research.

[§§] Corresponding authors.

In experiments, the proposed method outperforms other state-of-the-art methods and has abundant biological interpretation in explaining differences between breast cancer grades.

Keywords: Multiple kernel learning (MKL); feature selection; omics data integration; breast cancer grade; biological interpretation.

1. Introduction

Breast cancer is the most common cancer in women and a leading cause of cancer death worldwide. An estimated 1.67 million new cancer cases were diagnosed in 2012, and nearly 12% of the whole world new breast cancer cases were diagnosed in China.[1] Since the global burden of breast cancer exceeds all other cancers in women, and the incidence and mortality rates of breast cancer are increasing, early diagnosis of the cancer seems to be the most practical way to lower the harm. Current routine management of breast cancer relies on the availability of robust clinical or pathological prognostic, predictive factors to guide patient decision-making and the selection of potentially suitable treatment options. In early-stage breast cancer, where the use of systemic therapy has to be determined for every patient, the three main prognostic determinants used in routine practice are lymph node status, tumor size, and histological grade.[2] Within the last several decades, histologic grade has become widely accepted as a powerful indicator of early diagnosis and prognosis in breast cancer.

Histologic grade of cancer represents the morphological assessment of tumor biological characteristics and has been shown to be able to generate important information related to the clinical behavior of cancer such as cancer's malignancy and aggressiveness. A popular cancer grading system uses four levels of malignancy (G1–G4), reflecting the combined level of cell appearance abnormality, deviation in growth rate from the normal cells and the degree of invasiveness and dissemination. These pathological measures have been found to be in general concordance with the level of cellular differentiation by American Joint Commission on Cancer (AJCC).[3] More precisely, G1, G2, G3 and G4 are referred to as well-differentiated, moderately-differentiated, poorly-differentiated and undifferentiated conditions. Until now, there has not been a universal grading system for all cancers, however, different grading systems have been proposed for different cancers. The Nottingham (Elston–Ellis) modification of the Scarff–Bloom–Richardson grading system, also known as the Nottingham Grading System (NGS),[4] is the most well-known grading system recommended for breast cancer and based on assessments of nuclear grade, tubule formation, and mitotic rate.

Due to the significance of cancer grading, many researchers proposed cancer grade prediction methods based on diverse high-throughput omics data. Cui *et al.* identified a number of gene combinations whose expression patterns serve well as signatures of different grades based on gastric cancer data aiming to design grading score predictor.[5] Yao *et al.* presented several gene panels as signatures for distinguishing breast cancer grades and stages through studying on RNA-seq data.[6]

Shivang *et al.* integrated multi-scale image information and domain-specific information for breast and prostate grading score prediction.[7] Advances in different type of sequencing have revolutionized and facilitated the field of bioinformatics research, making it possible even for small research group to generate large amounts of sequencing data very rapidly and at a substantially lower cost. Moreover, what is the most important is that it is increasingly common to derive multiple types of omics data from the same patient. However, hardly most of researches tackle cancer grade prediction via integrating multiple omics data. Meanwhile, the emergence of a large number of high quality and confidence omics databases also provides us more possibility and convenience to collect these different types of data for the analysis that is referred to as integration analysis of heterogeneous data in data science terminology. Indeed, the emerging approaches for heterogeneous data integration analysis in biological field mainly include two paradigms: multi-staged and meta-dimensional analysis.[8] It is usually hard to fuse heterogeneous data, but the kernel methods can be applied to merge the inherent distinction among heterogeneous data and are hence suitable in this set-up.

Support vector machine (SVM) is known as a well-studied and classical machine learning method for classification problem. However, many real data are linearly inseparable, the kernel hence plays a prominent role in SVM. One of the fine properties of kernel methods is that we can map low-dimensional features to a high-dimensional even infinite-dimensional space, where data are linear separable. However, only using single kernel restricts the ability of SVM in integrating multiple types of omics data simultaneously. On the contrary, multiple kernel learning (MKL) that adopts the second strategy based on data level, encodes data of different types into kernels respectively as input for its own model, is exactly right applied for heterogeneous data. MKL aims to simultaneously learn a combined kernel and the associated predictor in supervised learning settings. There are two common methods to combine different kernels: linear method and nonlinear method. Among them, the most state-of-the-art method is linear MKL using weighted sum of different kernels as a combined kernel.[9]

In this paper, we proposed an improved supervised method based on linear MKL framework fusing heterogeneous omics data of breast cancer from the Cancer Genome Atlas (TCGA) (https://tcga-data.nci.nih.gov/tcga/) and biological pathway information from Kyoto Encyclopedia of Genes and Genomes (KEGG)[10] to predict grade scores of breast cancer and evaluate the significance of various pathways between different breast cancer grades. The workflow of the proposed method, which contains two parts of pipeline, is illustrated in Fig. 1 as below. In the first part, we focused on MKL fusing heterogeneous omics data aiming at predicting the grade score of breast cancer. (1) We collected different omics data of breast cancer, including both gene expression data and methylation data, which are derived from the same patient and annotated with grade score. (2) We preprocessed these data by purging, normalizing and gene mapping. Then, we performed gene feature selection

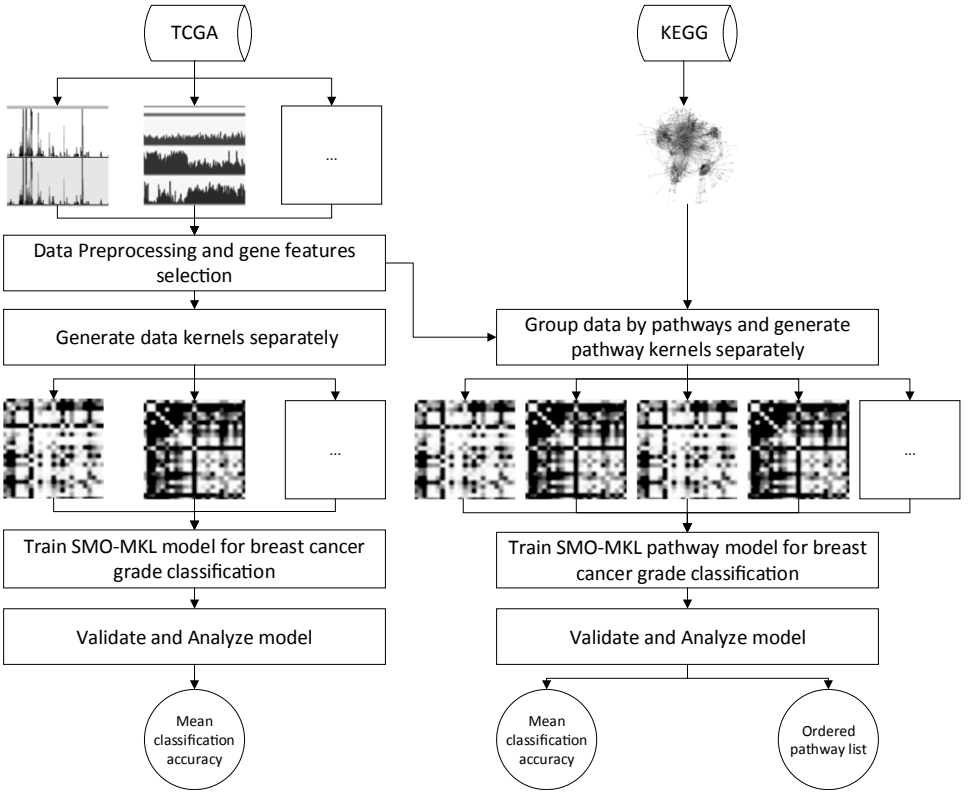


Fig. 1. The workflow of working model.

method based on the intrinsic characteristics that determine their relevance or discriminant powers with regard to the target grade to different omics data matrix separately. (3) We generated kernels utilizing data matrices as input for subsequent model, and then trained linear MKL model to obtain the classification result of the predictor. Then, we used the mean accuracy of classification to indicate the efficiency of model, and compared the results of proposed method with other traditional methods. In the second part, we did the same process as (1) and (2) in the first part. (3) Then we grouped data matrix mentioned in the previous step into sub-matrices, each of these rows representing genes belong to the same biological pathway respectively, and generated kernels utilizing these sub-matrices as input for subsequent model. (4) We trained a linear MKL model to acquire the weights of various pathways between different breast cancer grades. Furthermore, our method involving biological pathway information is capable of providing us abundant biological interpretation of the differences among various grades, which merits deeper exploration.

2. Methods

2.1. *Data preprocessing*

In this paper, we tried to predict breast cancer grade scores using MKL with breast cancer data from TCGA, which cover 1098 breast cancer patients with various clinical descriptions and omics data. Indeed, we attempted to collect as many samples with both clinical descriptions we concerned and distinct omics data as possible, which empower our method to improve the robustness and reliability in deciphering the differences among breast cancer grades. In this study, we filtered 610 samples, each of which measured expression data, methylation data and grade score simultaneously, can maximize our demand. In these 610 cancer samples, 75 samples are well-differentiated (G1), 289 samples are moderately-differentiated (G2) and 246 samples are poorly differentiated (G3). The expression data were obtained from the level 3 data of RNASeqV2, and the methylation data were extracted from the level 3 data of DNA Methylation 450k. Moreover, in order to unify the format of different omics data and facilitate following gene feature selection. As to expression data, we firstly converted the ID of probes to gene symbols, and then merged corresponding values based on mapping between probes and genes. As to methylation data, we just considered the relationship of regulation between the transcription of gene and the hyper-methylation or hypo-methylation of corresponding promoter. Further, we used this relationship to map probes of methylation data to gene symbols, meanwhile, aggregated associated Beta values of probes reflecting the activities of promoters into a single value reflecting the regulation of corresponding gene. Eventually, each omics data is represented as a two-dimensional matrix with each row representing one gene symbol, each column representing one sample with a grade score.

2.2. *Feature selection*

Since a majority of genes may contribute a little to classification results, we applied a feature selection method to select genes. Generally, there are three types of feature selection methods: filter methods, wrapper methods and embedded methods.[11] In this paper, an advanced filter method was employed to select informative genes by relevant statistical score of each gene and presents an appointed threshold by which informative genes are selected. The reason why we used the filter method is that we can independently select as many informative genes as possible without concerning about the existence of consistent patterns among different genes. Despite maybe involving redundant genes, we can indeed obtain more abundant biological interpretations in our following pathway analysis. In particularly, first, we performed Wilcoxon rank-sum test on each omics data to obtain the p-values of genes, respectively. Then, we adjusted these p-values through Benjamini-Hochberg False Discovery Rate (BH-FDR) controlling procedure and attained corresponding q-values for these genes. Moreover, we set a threshold of q-values, which is a fraction of discoveries and is tolerated to be false, to filter relatively significant genes. In this

study, we chose the genes with acceptable FDRs below 0.05 as informative genes. Finally, to further eliminate biases yielded by noisy omics data, such as genes with higher significances but smaller values in omics data. Thus, we measured genes selected above through calculating their fold change (FC) levels between any two grades, and obtained down- or up-regulated genes by whether their FCs are less than threshold 0.5 or greater than threshold 2.

2.3. *MKL kernels generating*

For breast cancer grade prediction, we constructed kernels of MKL based on gene expression data and methylation data, respectively. It seems that omics data of different types have different intrinsic scales, so we need to normalize these kernels utilizing the formula below:

$$K_{\text{norm}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i)K(\mathbf{x}_j, \mathbf{x}_j)}}$$

where K represents kernel function, $\mathbf{x}_i$ represents the ith data point and $\mathbf{x}_j$ represents the jth data point.

In this paper, three types of kernel functions including Linear kernel, Gaussian kernel and Polynomial kernel were employed. Linear kernel is the most commonly used in high dimensional space because the data with a large number of features tend to be linearly separable, meanwhile, Gaussian kernel and Polynomial kernel are often used in handling nonlinear separable problem.

To incorporate multiple omics data with the pathway information based on our method for following pathway analysis on breast cancer grade, we split both gene expression data matrix and methylation data matrix into more delicate matrices based on gene sets derived from various pathways. It means that we can construct pathway-based matrices, each rows of which represents a fraction of significant genes falling into identified pathway, as well, each columns of which represents all samples we collected just like expression data or methylation data. Therefore, we can generate kernels for each pathway as input to the proposed model, in which each kernel is assigned a coefficient indicating the contribution to the breast cancer grade prediction. Eventually, we explored the connections between breast cancer grade and the most significant pathways retrieved from the proposed model.

2.4. *SMO-MKL model*

2.4.1. *The MKL formulation*

In this paper, we focused on the mission where kernels encoding different omics data are learnt to be an optimal linear combination of given base kernels with non-negative weights to boost our final model. This mission actually aims to solve a standard linear MKL problem, and the original dual problem of linear MKL is

formulized as follows:

$$\begin{aligned}
&\min_{\mathbf{d}} \max_{\boldsymbol{\alpha}} \left(\mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2} (\boldsymbol{\alpha} \circ \mathbf{y})^T K(\mathbf{d}) (\boldsymbol{\alpha} \circ \mathbf{y}) \right). \\
&K(\mathbf{d}) = \sum_{\ell=1}^{s} d_\ell \mathbf{K}_\ell, \ell \in [1, s]. \\
&\text{s.t. } 0 \leq \alpha_i \leq C, \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m] \quad \text{and} \quad \mathbf{d} \geq 0,
\end{aligned} \tag{1}$$

where $\mathbf{K}(\mathbf{d})$ is a function with respect to coefficient vector $\mathbf{d}$, which weigh the significance of different kernels and indicate the relative importance of different kernels to learning task. m is the number of data points, and $\mathbf{y}$ represents data labels. C is the cost for misclassification and α are Lagrange multipliers directly related to the model that we need to optimize. Actually, the standard MKL formulation, which learns a linear combination of base kernels subject to l_1 regularization, leads to a dual which is not differentiable, and state of the art algorithms today overcome this limitation by solving an intermediate saddle point problem rather than the dual itself.[12,13]

However, Vishwanathan *et al.* demonstrated that linear MKL regularized with the p-norm squared, or with certain Bregman divergences can indeed be trained using more efficient process. The proposed algorithm retains both simplicity and efficiency and is significantly faster than the specialized p-norm MKL solvers.[14] Now, our new dual objective, unlike the objective in formula (1), is differentiable with respect to α. The p-norm MKL problem as formula (2) is described in the below.

$$\begin{aligned}
&\min_{\mathbf{d}} \max_{\boldsymbol{\alpha}} \left(\mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2} (\boldsymbol{\alpha} \circ \mathbf{y})^{\mathrm{T}} K(\mathbf{d}) (\boldsymbol{\alpha} \circ \mathbf{y}) + \frac{\lambda}{2} \left(\sum_{\ell=1}^{s} d_\ell^p \right)^{\frac{2}{p}} \right). \\
&\text{s.t. } 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m] \quad \text{and} \quad \mathbf{d} \geq 0.
\end{aligned} \tag{2}$$

2.4.2. *SMO optimizing*

Sequential minimization optimization (SMO), which can be used for training SVM, is a very efficient optimization process. The dual problem of SVM is a typical quadratic programming (QP) optimization problem and can be formulized as follows:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i \cdot \mathbf{x}_j). \\
&\text{s.t. } 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].
\end{aligned} \tag{3}$$

where $\mathbf{x}_i$ represents the ith data point, and y represents the ith data label. C is the cost for misclassification and $\boldsymbol{\alpha}_i$ is Lagrange multiplier directly related to the model

to be optimized. And, k is the kernel function that projects the data into a high dimension.

Instead of solving a large QP problem, which is both time and memory consuming, the SMO decomposes the large problem into a series of smallest possible QP problems each involving only two Lagrange multipliers, because all Lagrange multipliers must obey a linear equality constraint. There are only two common components in SMO: an analytic method for solving the two Lagrange multipliers, and a heuristic for choosing which two multipliers to joint optimize. In a word, the SMO algorithm can tolerate where two variables are selected and optimized using gradient methods and iterative process until convergence.

The SMO is simple, easy to implement and adapt, and efficiently scales to large problems, so it has gained widespread acceptance. Hence, training using SMO has been a long-standing goal in MKL for the very same reasons.

2.4.3. *Using SMO-MKL algorithm*

The dual of standard p-norm MKL is differentiable, so it can be optimized using SMO style co-ordinate ascent directly. The standard SMO-MKL problem as formula (3) can be described in the below.

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \left(\mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{8\lambda}\left(\sum_{\ell=1}^{s} \left((\boldsymbol{\alpha}\circ\mathbf{y})^{\mathrm{T}} K_\ell (\boldsymbol{\alpha}\circ\mathbf{y}) \right)^q \right)^{\frac{2}{q}} \right). \\
&d_\ell = \frac{1}{2\lambda}\left(\sum_{\ell=1}^{s} (\boldsymbol{\alpha}\circ\mathbf{y})^{\mathrm{T}} K_\ell (\boldsymbol{\alpha}\circ\mathbf{y}) \right)^{\frac{1}{q}-\frac{1}{p}} ((\boldsymbol{\alpha}\circ\mathbf{y})^{\mathrm{T}} K_\ell (\boldsymbol{\alpha}\circ\mathbf{y}))^{\frac{q}{p}}. \qquad (4) \\
&\text{s.t.} \quad 0 \le \alpha_i \le C, \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1,m] \quad \text{and} \quad \frac{1}{p} + \frac{1}{q} = 1.
\end{aligned}$$

where p is the regularization norm and q is a parameter simplifying the objective.

The SMO-MKL uses larger p-norm regularization leading to a dense solution on kernel weight, which produces a better result than that from the MKL using 1-norm regularization which leads to a sparse solution.[12,15–18] Dense solution also provides us a better interpretation of coefficients corresponding to different kernels. This is the most important reason why we chose the SMO-MKL. We implemented the SMO-MKL algorithm based on LibSVM code.[19]

3. Experimental Results

3.1. *The performance of proposed model*

3.1.1. *Using omics data of different types*

In this section, we integrated different omics data after preprocessing and feature selecting, with which we implemented the SMO-MKL model to solve classification problem of any two grades. We performed our model on all pairwise grades, including

Table 1. The accuracies of all pairwise grades with different combination of data and kernels.

Grade pairs	G1 versus G2	G2 versus G3	G1 versus G3
SVM^l-Expression Data	$76.0 \pm 2.7\%$	$71.4 \pm 4.8\%$	$87.2 \pm 4.4\%$
SVM^l-Methylation Data	$76.3 \pm 1.9\%$	$75.2 \pm 4.7\%$	$87.0 \pm 5.0\%$
SVM^p-Expression Data	$76.0 \pm 2.7\%$	$73.8 \pm 6.1\%$	$87.2 \pm 4.6\%$
SVM^p-Methylation Data	$76.3 \pm 3.0\%$	$75.1 \pm 6.7\%$	$85.1 \pm 3.6\%$
SVM^g-Expression Data	$72.7 \pm 5.1\%$	$75.3 \pm 7.1\%$	$82.2 \pm 7.2\%$
SVM^g-Methylation Data	$76.6 \pm 2.2\%$	$70.5 \pm 6.1\%$	$87.0 \pm 3.7\%$
MKL-Expression Data	$76.3 \pm 4.9\%$	$78.6 \pm 6.9\%$	$88.2 \pm 5.7\%$
MKL-Methylation Data	$76.6 \pm 4.2\%$	$75.5 \pm 6.2\%$	$87.8 \pm 4.5\%$
MKL- Multiple Data	$\mathbf{77.4 \pm 4.1\%}$	$\mathbf{79.7 \pm 6.7\%}$	$\mathbf{89.1 \pm 3.1\%}$

G1 versus G2, G2 versus G3 and G1 versus G3. To validate the performance of the proposed model, we calculated the average accuracies and the area under the curves (AUCs) by 10-fold cross-validation both based on MKL models and SVM models encoding different kernels with expression data alone, methylation data alone and all data respectively. Table 1 shows our method achieved the best accuracies in all pairwise grades, which suggests our method is connected with high precision in predicating grades, as well, Table 2 shows our method achieved the best AUCs in all pairwise grades, which indicates our method is endowed with strong robustness in handling skewed data. Also, we measured the average sensitivities and specificities using the same settings above, see Tables S1 and S2 respectively.

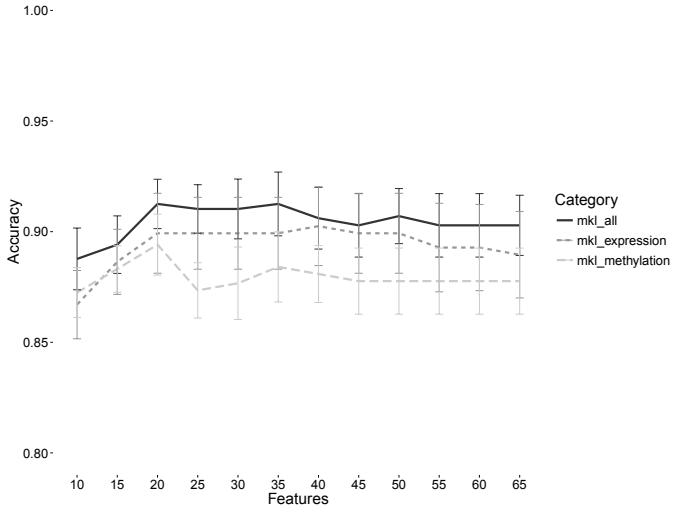
The experimental results of the average accuracies in classifying G1 versus G2, G2 versus G3 and G1 versus G3 using expression and methylation data are $77.4 \pm 4.1\%$, $79.7 \pm 6.7\%$ and $89.1 \pm 3.1\%$, respectively. The experimental results of the AUCs in classifying G1 versus G2, G2 versus G3 and G1 versus G3 using expression and methylation data are 0.685, 0.861 and 0.959, respectively. It can be found that the results produced by multiple data with multiple kernels outperform those using individual data with one specific kernel or individual data combining with multiple kernels in all pairwise grades. However, the results of classifying both G1 versus G2 and G2 versus G3 are slightly unsatisfactory, which also imply that expression data

Table 2. The AUCs of all pairwise grades with different combination of data and kernels.

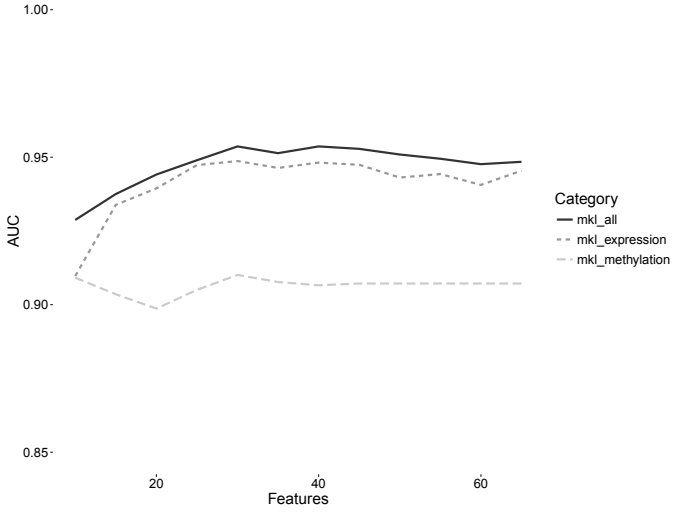
Grade pairs	G1 versus G2	G2 versus G3	G1 versus G3
SVM^l-Expression Data	0.535	0.762	0.919
SVM^l-Methylation Data	0.564	0.801	0.916
SVM^p-Expression Data	0.660	0.814	0.891
SVM^p-Methylation Data	0.572	0.828	0.896
SVM^g-Expression Data	0.572	0.812	0.869
SVM^g-Methylation Data	0.578	0.745	0.876
MKL-Expression Data	0.606	0.840	0.926
MKL-Methylation Data	0.554	0.826	0.908
MKL- Multiple Data	**0.685**	**0.861**	**0.959**

and methylation data might be low-resolution to characterize the differences among grades. The similar results yielded by proteomic data have been reported that the boundaries between G1 versus G2 and G2 versus G3 are also somewhat obscure.[20]

We also measured the performance of the proposed model on different selected genes. We constructed kernels with combining the top 10 to top 65 genes sorted by



(a)



(b)

Fig. 2. The accuracies (a) and the AUCs (b) of the proposed model based on MKL with different selected genes using expression data, methylation data and both of them, respectively.

q-values in ascending order from expression data and methylation data. The average accuracies and the average AUCs of proposed model using different selected genes are given in Fig. 2. Apparently, it can be observed that the number of selected genes has a certain influence on accuracy in some extent. We achieved the best mean accuracy $91.6 \pm 3.8\%$ by combining the top 20 genes from expression data and methylation data. The overlap of genes selected by each iteration of 10-folds cross-validation as well as their annotations are attached to supplement materials, see Table S3. Also we calculated the sensitivities and specificities using the same setting mentioned above, see Figs. 1 and 2.

3.1.2. *Comparison with other approaches*

To further demonstrate the performance of the proposed model, we compared it with two state-of-the-art approaches, random forest and ensemble. The random forest approach was implemented using the randomForest R-package,[21] which had been modified to handle multiple data sources. The ensemble approach was implemented using the caretEnsemble R-package,[22] which had been modified to work with multiple data sources via the greedy stepwise approach. In this study, we applied the same feature selection procedures in the proposed method to these approaches, and attained the average accuracies by 10-fold cross-validation of these two approaches based on expression data and methylation data. The results of comparison among various algorithms are shown in Fig. 3. The average accuracies of random forest approach are $69.5 \pm 8.5\%$, $72.6 \pm 4.9\%$ and $86.6 \pm 7.4\%$ in classifying G1 versus G2, G2 versus G3 and G1 versus G3, respectively. While the average accuracies of ensemble method are $69.5 \pm 5.3\%$, $58.2 \pm 5.6\%$ and $82.5 \pm 5.9\%$ in classifying G1 versus G2, G2 versus G3 and G1 versus G3, respectively. Moreover, the consistent results of AUCs are shown in Fig. 3. Obviously, we achieved the best results in all pairwise grades by using the proposed MKL model. Also we calculated sensitivities and specificities using the same settings above for all pairwise groups, see Figs. S3 and S4.

3.2. *Biological pathways analysis*

For purpose of enriching biological interpretation in depicting the distinctions residing among grades, we divided both gene expression data matrix and methylation data matrix into more delicate matrices by gene sets derived from various pathways, and then used these matrices to generate corresponding pathway-based kernels. Here, the value of each coefficient λ of kernel assigned by the proposed model implies how important the influence of associated pathway is in distinguishing breast cancer grades. We sorted 186 pathways from KEGG based on relevant kernel coefficients, and captured the pathways with coefficients greater than 1. The selected pathways are primary related to cell differentiation, cell development and several pathways, which have been found to have a strong relationship with breast cancer, such as sphingolipid metabolism, hedgehog signaling, ECM receptor interaction, etc.

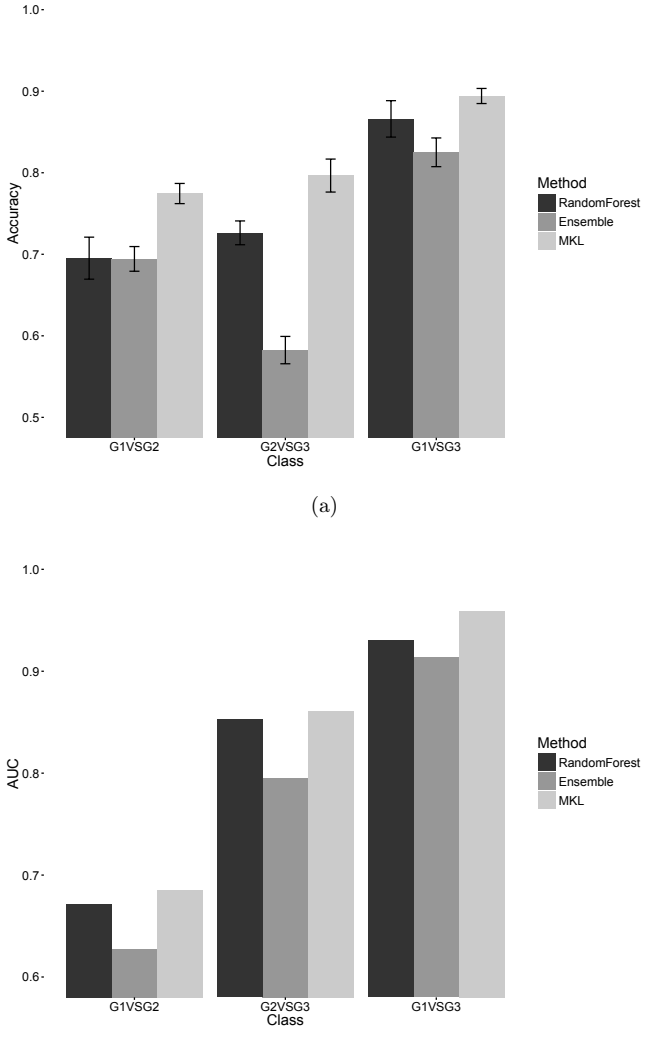


Fig. 3. The average accuracies (a) and the AUCs (b) of random forest, ensemble method and the proposed method based on MKL in all pairwise grades.

The top 10 significant pathways are shown in Table 3, and the complete pathways are listed in supplement materials as shown in Table S4.

3.2.1. *Metabolism pathways*

Obviously, most of the pathways with significant classification impact are related to metabolism as shown in Tables 3 and S4. We can see that sphingolipid metabolism pathway owns the highest scores. Sphingolipid metabolism pathway is activated during normal cell growth and division not only to provide a basic supply of

Table 3. The top 10 significant pathways retrieved from pathway-based MKL results.

WEIGHT	PATHWAY
5.697	KEGG_SPHINGOLIPID_METABOLISM
5.137	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_KERATAN_SULFATE
4.853	KEGG_HEDGEHOG_SIGNALING_PATHWAY
4.795	KEGG_GLUTATHIONE_METABOLISM
4.715	KEGG_PEROXISOME
4.130	KEGG_PPAR_SIGNALING_PATHWAY
3.755	KEGG_FATTY_ACID_METABOLISM
3.684	KEGG_ECM_RECEPTOR_INTERACTION
3.412	KEGG_PYRUVATE_METABOLISM
2.524	KEGG_TAURINE_AND_HYPOTAURINE_METABOLISM

structural and functional metabolites but also to serve as second messengers.[23,24] Sphingolipids contain a complex family of naturally occurring molecules that are enriched in cellular membranes and serve as a reservoir for the production of bioactive metabolites, including sphingosine, ceramide, sphingosine-1-phosphate and ceramide-1-phosphate. Moreover, sphingolipids involved in the regulation of cancer cell proliferation, survival and migration, specifically sphingosine-1-phosphate, have been shown to mediate numerous pro-oncogenic processes including evasion of apoptosis cell transformation, uncontrolled proliferation, desensitization of anti-growth agents, angiogenesis and metastasis.[25]

Glycosaminoglycan biosynthesis keratan sulfate pathway is also worth noting. The cell surface-associated glycan structure plays a key role in controlling cell behavior in extracellular matrix (ECM) environment during both physiological and pathological processes by regulating cell to ECM and cell to cell interactions. Further, proteoglycans consisting of linear glycosaminoglycan chains are a family of glycans expressed on the cell surface, cytoplasm, and extracellular matrix. These proteoglycans can regulate cell migration, invasion and growth of various cancer cells. Moreover, keratan sulfate is any of several sulfated glycosaminoglycan, and it has been reported having a strong relationship with breast cancer.[26,27] The expression of several core proteins of keratan sulfate, including lumican and decorin, has been found to differ in neoplastic tissue in comparison to adjacent tissue. Lumican was found to be significantly up-regulated while decorin was lowly expressed in the malignant tissue.[28] On the other hand, both of these two core proteins of keratan were demonstrated having lower expression.[29] However, the implication of role of keratan sulfate is still under investigation.

Moreover, taurine and hypotaurine metabolism pathway needs to be considered. Taurine, the most abundant free amino acid in humans, has numerous potential health benefits through its antioxidant and ant-inflammatory properties, while taurine is reported to be downregulated in cancer.[30] A large number of studies and clinical applications showed that taurine as an effective antioxidant may hinder an increase in reactive oxygen species in tumors, where elevated reactive oxygen species may be a driver for initiation and early development of breast cancer,[31] thereby

preventing the development of cancer. And taurine promoted apoptosis in breast cancer and repressed the growth of tumor by regulating the expression of apoptosis-related proteins of mitochondria.[32]

Glutathione (GSH) metabolism pathway also should be stressed. In many normal and malignant cells, increased GSH level is associated with a proliferative response and is essential for cell cycle progression. A key mechanism for GSH's role in DNA synthesis associated with the maintenance of reduced glutaredoxin or thioredoxin which is required for the activity of ribonucleotide reductase, the rate-limiting enzyme in DNA synthesis. As well, a high percentage of tumor cells with high GSH content were able to survive in the presence of the nitrosative and oxidative stress, thereby representing the main task force in the metastatic invasion.[33]

Furthermore, many common metabolic pathways in cancer are highlighted, such as purine metabolism pathway, pyruvate metabolism pathway, fatty acid metabolism pathway, starch and sucrose metabolism pathway have also been reported to be strongly associated with breast cancer cell development and growth.

3.2.2. *Signaling pathways*

Another category of pathways with significant impact in cancer grading is signaling pathway. Hedgehog signaling pathway, which transmits information such as time and position dependent expression patterns to embryonic cells required for proper development, plays a crucial role in vertebrate embryogenesis by controlling cell fate, patterning, proliferation, survival and differentiation, and thus affects development from embryonic stage.[34] In the adult organism, Hedgehog signaling remains active and is involved in the regulation of tissue homeostasis, regeneration and stem cell maintenance.[35] Aberrant activation of the Hedgehog signaling pathway has been unambiguously tied to breast cancer development and progression.[36,37]

Peroxisome proliferator-activated receptors (PPARs) to which PPAR signaling pathway related are nuclear hormone receptors that are activated by fatty acids and their derivatives, and play essential roles in the regulation of cellular differentiation, development, immune response and tumorigenesis of higher organisms.[38,39] Mitogen-activated protein kinase (MAPK) cascade to which MAPK signaling pathway related is a highly conserved module, and is involved in various cellular functions including cell proliferation, differentiation and migration.[40,41] And calcium signaling pathway plays a role in processes important in cancer, such as proliferation, apoptosis,[42,43] invasion and migration,[44,45] and is a key in cell signaling pathways in various differentiation stages of stem cells.[46] Chemokine signaling pathway regulates plethora of biological processes of hematopoietic cells to lead cellular activation, differentiation and survival.[47] Toll-like receptor signaling pathway, a quintessential signaling mechanism of inflammation uses the Toll-like receptor (TLR) family, where TLR family is a highly conserved family of transmembrane proteins, recognizes a range of microbial agents as well as endogenous macromolecules released by injured tissue. Notably, inflammation has emerged as a non-mutational driver of tumor

development and progression, and has been associated with higher tumor grades and a poor prognosis.[48]

3.2.3. *Regulatory pathways*

The other significant pathways are those about regulatory. Extracellular matrix (ECM) to which ECM receptor interaction pathway related, consists of a complex network of macromolecules with distinctive physical, biochemical, and biomechanical properties secreted by cells and serves an important role in tissue and organ morphogenesis and in the maintenance of cell and tissue structure and function. Additionally, ECM indirectly affects cancer cells by promoting formation of a tumor microenvironment. Deregulation of ECM dynamics can facilitate cellular dedifferentiation and cancer stem cell expansion, and directly lead to cellular transformation and metastasis.[49] And most commonly, stromal cells, endothelial cells, immune cells, and fibroblasts, are the main initial culprits that cause abnormal ECM production.[50–52] And cytokines to which cytokine–cytokine receptor interaction related are soluble extracellular proteins or glycoproteins, and they are crucial intercellular regulators and mobilizers of cells engaged in cell growth and differentiation.[53] As well cytokines are intricately involved in all immune reactions, and malignant tumor utilizes these immune reactions within its microenvironment to prevent activation of immunological effector functions, thereby protecting the tumor from a potential immune attack.[54]

4. Discussion

With the development of high-throughput technology, the researchers can acquire a large number of omics data with different types from several public databases. Histologic grade of cancer is related to tumor biological and morphological characteristics. In this paper, we proposed a method for cancer grade prediction and pathway analysis on multiple omics data using MKL algorithm. The proposed method has satisfactory performance through comparison with other state-of-the-art approaches in analysis of breast cancer grade. The reason of satisfactory performance is that each kernel has an adaptive weight for individual kernel and the method maximally utilizes the information encoded in each kernel according to different problem. In addition, we can use MKL algorithm for biological pathway analysis by the weight of each pathway, which indicates the relative significance. Using the method, we found that several pathways have strong association with characteristics of breast cancer grade, such as cell differentiation, growth, proliferation and dissemination.

We also predicted cancer grade by using MKL model based on biological pathway information. We observed the involvement of biological pathway information leading to a little decline somewhat in average accuracy: the disadvantage of 2- or p-norm MKL is that 2- or p-norm MKL cannot eliminate weak kernels as 1-norm MKL does when there indeed exists noisy in our raw data. In other words, 2- or p-norm MKL is less robust than other approaches. Owing to this disadvantage, we can explain why average accuracy will decline when more kernels involved in MKL,[55,56] but it still can be tolerated.

Owing to the cost of computation, in this paper, we just chose a set of empirical parameters as optimal. Actually, we set gamma to 0.01, degree to 2 and intercept to 0 respectively when using polynomial kernel, set gamma to 0.01 when using Gaussian kernel. It is reasonable to expect that we can promote average accuracy beyond 90% through rationally tuning parameters. With heuristic search or grid search, optimal parameter can be acquired for the proposed model. Meanwhile, more various types of data such as copy number variation data, mutation data, micro-RNA data, proteomic data and even metabolic data also have potential to enhance the performance of our model. And for the methodology, there indeed exists a large number of MKL implementations[57] that strive to promote the efficiency and performance, such as nonlinear kernel combination[16,58] and sample adaptive parameter that can denoise automatically.[59] All these refinements will provide us more possibility to further pave the way for improved breast histological grading and prognostication in the future.

More than pathway information, several biological information, such as gene ontology information, gene co-expression information and domain knowledge with respect to specific gene sets, can be involved in further biological function analysis. The generated kernel coefficient indicates the relative significance of each kernel constructed by using a specific biological function individually. And then all these coefficients can further be used to give a rational explanation to the contribution of enriched biological functions to our grade classification problem. Therefore, we are able to gain new insight into the differences between cancer grades in biological function level, and explore the mechanism hided in phenotype of each cancer grade through biological experiments under the guidance.

Finally, we simply summarized molecular subtypes and grades of samples, see Tables S5 and S6. Interestingly, we found our misclassified samples mainly concentrate on two subtypes: LumA and LumB, with which samples mainly fall into G1 and G2. While it seems that other subtypes such as Basal and Her2 may have distinct boundaries and are seldom misclassified, samples with these subtypes mainly fall into G3. We will further design more delicate classifiers focusing on grade classification of breast cancer in different subtypes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61472158, 61402194, 61572227, 61572228), China Postdoctoral Science Foundation (2014T70291) and Development Project of Jilin Province of China (20140101180JC).

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F, Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012, *Int. J. Cancer* **136**:E359–E386, 2015.

2. Rakha E, Reis-Filho J, Baehner F, Dabbs D, Decker T, Eusebi V, Fox S, Ichihara S, Jacquemier, J, Lakhani S, Palacios J, Richardson A, Schnitt S, Schmitt F, Tan P-H, Tse G, Badve S, Ellis I, Breast cancer prognostic classification in the molecular era: The role of histological grade, *Breast Cancer Res.* **12**: 207, 2010.
3. Edge SB, Compton CC, The American Joint Committee on Cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM, *Ann. Surg Oncol*, **17**:1471–1474, 2010.
4. Elston CW, Ellis IO, Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up, *Histopathology* **19**:403–410, 1991.
5. Cui J, Li F, Wang G, Fang X, Puett JD, Xu Y, Gene-expression signatures can distinguish gastric cancer grades and stages, *PLoS One* **6**, 2011.
6. Yao F, Zhang C, Du W, Liu C, Xu Y, Identification of gene-expression signatures and protein markers for breast cancer grading and staging, *PLoS One* **10**, 2015.
7. Naik S, Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J, Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology, Biomedical Imaging: From Nano to Macro, 2008, *5th IEEE Int Symp on ISBI*, pp. 284–287, 2008.
8. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D, Methods of integrating data to uncover genotype-phenotype interactions, *Nat Rev Genet* **16**:85–97, 2015.
9. Gönen M, Alpaydın E, Multiple kernel learning algorithms, *J Mach Learn Res* **12**:2211–2268, 2011.
10. Kanehisa M, Goto S, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**:27–30, 2000.
11. Tang J, Alelyani S, Liu H, *Feature Selection for Classification: A Review, Data Classification: Algorithms and Applications*, (2014) 37.
12. Kloft M, Brefeld U, Laskov P, Müller K-R, Zien A, Sonnenburg S, Efficient and accurate lp-norm multiple kernel learning, *Adv Neural Inf Process Syst*, 2009, pp. 997–1005.
13. Rakotomamonjy A, Bach F, Canu S, Grandvalet Y, Simple MKL, *J Mach Learn Res* **9**;2491–2521, 2008.
14. Sun Z, Ampornpunt N, Varma M, Vishwanathan S, Multiple kernel learning and the SMO algorithm, *Adv. Neural Inf Process Syst* 2010, 2361–2369.
15. Cortes C, Mohri M, Rostamizadeh A, L 2 regularization for learning kernels, *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2009, pp. 109–116.
16. Cortes C, Mohri M, Rostamizadeh A, Learning non-linear combinations of kernels, *Adv Neural Inf Process Syst* 2009, 396–404.
17. Varma M, Babu BR, More generality in efficient multiple kernel learning, *Proc the 26th Annual Int Conf on Machine Learning*, ACM, pp. 1065–1072, 2009.
18. Kloft M, Brefeld U, Sonnenburg S, Zien A, Lp-norm multiple kernel learning, *J Mach Learn Res,* **12**:953–997, 2011.
19. Chang C-C, Lin C-J, LIBSVM: A library for support vector machines, *ACM Trans Intell Syst Technol,* **2**:27, 2011.
20. Olsson N, Carlsson P, James P, Hansson K, Waldemarson S, Malmström P, Fernö M, Ryden L, Wingren C, Borrebaeck CAK, Grading breast cancer tissues using molecular portraits, *Mol Cell Proteomics* **12**:3612–3623, 2013.
21. Breiman L, Random forests, *Mach Learn,* **45**:5–32, 2001.
22. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A, Ensemble selection from libraries of models, *Proc 21st Int Conf Machine Learning* ACM, pp. 18, 2004.

23. Aoyagi T, Nagahashi M, Yamada A, Takabe K, The role of sphingosine-1-phosphate in breast cancer tumor-induced lymphangiogenesis, *Lymphatic Res Biol* **10**:97–106, 2012.
24. Lebman DA, Spiegel S, Thematic Review Series: Sphingolipids. Cross-talk at the crossroads of sphingosine-1-phosphate, growth factors, and cytokine signaling, *J Lipid Res* **49**:1388–1394, 2008.
25. Sukocheva O, Wadham C, Role of sphingolipids in oestrogen signalling in breast cancer cells: An update, *J Endocrinol*, **220**:R25–R35, 2014.
26. Afratis N, Gialeli C, Nikitovic D, Tsegenidis T, Karousou E, Theocharis AD, Pavao MS, Tzanakakis GN, Karamanos NK, Glycosaminoglycans: Key players in cancer cell biology and treatment, *FEBS J* **279**:1177–1197, 2012.
27. Iida J, Dorchak J, Clancy R, Slavik J, Ellsworth R, Katagiri Y, Pugacheva EN, van Kuppevelt TH, Mural RJ, Cutler ML, Shriver CD, Role for chondroitin sulfate glycosaminoglycan in NEDD9-mediated breast cancer cell growth, *Exper Cell Res* **330**:358–370, 2015.
28. Leygue E, Snell L, Dotzlaw H, Troup S, Hiller-Hitchcock T, Murphy LC, Roughley PJ, Watson PH, Lumican and decorin are differentially expressed in human breast carcinoma, *J Pathol* **192**:313–320, 2000.
29. Troup S, Njue C, Kliewer EV, Parisien M, Roskelley C, Chakravarti S, Roughley PJ, Murphy LC, Watson PH, Reduced expression of the small leucine-rich proteoglycans, lumican, and decorin is associated with poor outcome in node-negative invasive breast cancer, *Clinical Cancer Res an Official Journal of the American Association for Cancer Research*, **9**:207–214, 2003.
30. Fong MY, McDunn J, Kakar SS, Identification of metabolites in the normal ovary and their transformation in primary and metastatic ovarian cancer, *PLoS ONE* **6**:e19963, 2011.
31. Zhang C, Cao S, Toole BP, Xu Y, Cancer may be a pathway to cell survival under persistent hypoxia and elevated ROS: A model for solid-cancer initiation and early development, *Int J Cancer* **136**:2001–2011, 2015.
32. Zhang X, Lu H, Wang Y, Liu C, Zhu W, Zheng S, Wan F, Taurine induces the apoptosis of breast cancer cells by regulating apoptosis-related proteins of mitochondria, *Int J Molecular Med* **35**:218–226, 2015.
33. Traverso N, Ricciarelli R, Nitti M, Marengo B, Furfaro AL, Pronzato MA, Marinari UM, Domenicotti C, Role of glutathione in cancer progression and chemoresistance, *Oxidat Med Cellular Longevity* (2013).
34. Barakat MT, Humke EW, Scott MP, Learning from Jekyll to control Hyde: Hedgehog signaling in development and cancer, *Trends Molecul Med* **16**:337–348, 2010.
35. Hooper JE, Scott MP, Communicating with Hedgehogs, *Nature Rev: Molecul Cell Biol* **6**:306–317, 2005.
36. Hui M, Cazet A, Nair R, Watkins D, O'Toole S, Swarbrick A, The Hedgehog signalling pathway in breast development, carcinogenesis and cancer therapy, *Breast Cancer Res* **15**:203, 2013.
37. Onishi H, Katano M, Hedgehog signaling pathway as a therapeutic target in various types of cancer, *Cancer Sci* **102**:1756–1760, 2011.
38. Vamecq J, Colet J-M, Vanden Eynde JJ, Briand G, Porchet N, Rocchi S, PPARs: Interference with Warburg'effect and clinical anticancer trials, *PPAR Res* (2012).
39. Tachibana K, Yamasaki D, Ishimoto K, Doi T, The Role of PPARs in Cancer, *PPAR Res* (2008).
40. Imajo M, Tsuchiya Y, Nishida E, Regulatory mechanisms and functions of MAP kinase signaling pathways, *IUBMB Life* **58**:312–317, 2006.

41. Wong KK, Recent developments in anti-cancer agents targeting the Ras/Raf/MEK/ERK pathway, *Recent Patents Anti-Cancer Drug Discovery* **4**:28–35, 2009.
42. Roderick HL, Cook SJ, Ca2+ signalling checkpoints in cancer: Remodelling Ca2+ for cancer cell proliferation and survival, *Nature Rev Cancer* **8**:361–375, 2008.
43. Monteith GR, Davis FM, Roberts-Thomson SJ, Calcium Channels and Pumps in Cancer: Changes and Consequences, *J Biol Chem* **287**:31666–31673, 2012.
44. Prevarskaya N, Skryma R, Shuba Y, Calcium in tumour metastasis: New roles for known actors, *Nature Rev Cancer* **11**:609–618, 2011.
45. Davis FM, Azimi I, Faville RA, Peters AA, Jalink K, Putney JW, Goodhill GJ, Thompson EW, Roberts-Thomson SJ, Monteith GR, Induction of epithelial-mesenchymal transition (EMT) in breast cancer cells is calcium signal dependent, *Oncogene* **33**:2307–2316, 2014.
46. Tonelli FM, Santos AK, Gomes DA, da Silva SL, Gomes KN, Ladeira LO, Resende RR, Stem cells and calcium signaling, *Adv Exper Med Biol* **740**:891–916, 2012.
47. Palacios-Arreola MI, Nava-Castro KE, Castro JI, García-Zepeda E, Carrero JC, Morales-Montor J, The role of chemokines in breast cancer pathology and its possible use as therapeutic targets, *J Immunol Res* **2014**, 2014.
48. Ridnour LA, Cheng RY, Switzer CH, Heinecke JL, Ambs S, Glynn S, Young HA, Trinchieri G, Wink DA, Molecular pathways: Toll-like receptors in the tumor microenvironment—poor prognosis or new therapeutic opportunity, *Clinical Cancer Res* **19**:1340–1346, 2013.
49. Lu P, Weaver VM, Werb Z, The extracellular matrix: A dynamic niche in cancer progression, *J Cell Biol* **196**:395–406, 2012.
50. Bhowmick NA, Neilson EG, Moses HL, Stromal fibroblasts in cancer initiation and progression, *Nature* **432**:332–337, 2004.
51. Orimo A, Gupta PB, Sgroi DC, Arenzana-Seisdedos F, Delaunay T, Naeem R, Carey VJ, Richardson AL, Weinberg RA, Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion, *Cell* **121**:335–348, 2015.
52. Quante M, Tu SP, Tomita H, Gonda T, Wang SS, Takashi S, Baik GH, Shibata W, DiPrete B, Betz KS, Bone marrow-derived myofibroblasts contribute to the mesenchymal stem cell niche and promote tumor growth, *Cancer Cell* **19**:257–272, 2011.
53. Dinarello CA, Historical review of cytokines, *Europ J Immunol* **37**:S34–S45, 2007.
54. Lippitz BE, Cytokine patterns in patients with cancer: A systematic review, *The Lancet Oncology* **14**:e218–e228, 2013.
55. Gehler P, Nowozin S, On feature combination for multiclass object classification, *IEEE 12th International Conference on Computer Vision* IEEE, pp. 221–228, 2009.
56. Cortes C, Mohri M, Rostamizadeh A, Generalization bounds for learning kernels, *Proc 27th Int Conf on Machine Learning (ICML-10)*, pp. 247–254, 2010.
57. Gönen M, Alpaydın E, Multiple kernel learning algorithms, *J Mach Learn Res* **12**:2211–2268, 2011.
58. Jain A, Vishwanathan SVN, Varma M, SPF-GMKL: Generalized multiple kernel learning with a million kernels, *Proc 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*, ACM, Beijing, China, pp. 750–758, 2012.
59. Liu X, Wang L, Zhang J, Yin J, Sample-adaptive multiple kernel learning, *28th AAAI Conf on Artificial Intelligence*, 2014.

Tianci Song is a Graduate Student in Computer Science at College of Computer Science and Technology, Jilin University (2014–now). He received B.S. degree in Computer Science from Jilin University, Changchun, in 2014. He was a visiting student in University of Georgia from 2015 to 2016, His research interests include machine learning methods, data mining, statistical modeling and bioinformatics.

Yan Wang Ph.D., graduated from the College of Computer Science and Technology of Jilin University, China, in 2007. He is currently an associate professor of the College of Computer Science and Technology of Jilin University. He was doing research on a bioinformatics collaborative project as a visiting scholar and post doctor at University of Georgia, USA from 2007 to 2011. During 2012 to 2013, he got a postdoctoral position in University of Trento, Italy. His research interests focus on Computational Intelligence and Bioinformatics, such as gene expression analysis, miRNA prediction, operon prediction, cancer marker prediction and protein–protein interaction network analysis. He has had over 50 research papers published including more than 20 indexed by SCI, such as *Bioinformatics*, *Plos One*, *Methods*, *BMC bioinformatics*, etc. And he has also presided over or taken part in several projects funded by the National Natural Science Foundation and "863" project in China, respectively.

Wei Du was born in Henan, China in 1983. He received the Ph.D. degree in Computer Science and Technology (Bioinformatics) from Jilin University, Changchun, China, in 2011. He is currently an associate professor in the College of Computer Science and Technology of Jilin University, China. He was a visiting scholar in University of Trento of Italy from 2010 to 2011, a visiting scholar in University of Georgia of USA from 2015 to 2016. He has published over 50 papers. His research was featured in *Bioinformatics, BMC bioinformatics, Artificial intelligence in medicine, Cognitive Computation, IEEE Transactions on NanoBioscience*, etc. His research interests include data Mining, machine learning, bioinformatics and system biology.

Sha Cao is a Ph.D. Student in Department of Statistics, University of Georgia (2011-now). She received B.S. degree in Mathematics from Beijing Normal University, Beijing, in 2011. Her research interests include: 1. The altered microenvironment for metastatic cancer cells that colonize a distant site change the metabolic patterns of these cells and drive their faster proliferation than localized cancer cells. 2. Discovering the differential functional patterns of large protein complexes and tight metabolic pathways in tumor versus normal tissues. 3. Fast proliferations of cancer cells can dilute the micro-environmental stresses in tumor tissues by measuring the relative number of cells in different cell cycle stages of a tumor tissue.

4. Epigenetic regulations can memorize cancer cells' survival under severe stresses by annotation of co-methylation modules.

Yuan Tian is a Ph.D. Student in Computer Science at College of Computer Science and Technology, Jilin University (2012-now). He received both B.S. (2008) and M.S. (2012) degree in Computer Science at College of Computer Science and Technology, Jilin University. His research interests include system biology, cancer bioinformatics and modeling cellular metabolism.

Yanchun Liang was born in Jilin, China in 1953. He received the Ph.D. degree in applied mathematics from Jilin University, Changchun, China, in 1997. He is currently a professor in the College of Computer Science and Technology of Jilin University, and the Department of Computer Science at Zhuhai College of Jilin University, China. He was a visiting scholar in Manchester University of U.K. from 1990 to 1991, a visiting professor in National University of Singapore from 2000 to 2001, a guest professor in Institute of High Performance Computing of Singapore from 2002 to 2004, a visiting professor in Trento University, Italy from 2006 to 2008, and a visiting professor in Missouri University, USA from 2011 to 2016. He has published over 400 papers. His research was featured in *IEEE Trans. on Knowledge and Data Engineering, IEEE Trans. on Systems, Man, and Cybernetics – Part A: Systems and Humans, IEEE Trans. on Geoscience and Remote Sensing, Bioinformatics, Journal of Micromechanics and Microengineering*, *Physical Review E*, *Smart Materials and Structures*, *Applied Artificial Intelligence*, etc. He was the recipient of several grants from NSFC, EU, etc. His research interests include computational intelligence, machine learning methods, text mining, MEMS modeling and bioinformatics.