Gene selection for cancer classification using the combination of SVM-RFE and GA

Xiaobo Li[*]

Department of Computer Science and Technology, College of Engineering, Lishui University, Lishui 323000, China

Received 1 March 2014, www.tsi.lv

Abstract

Gene selection is a key research issue in molecular cancer classification and identification of cancer biomarkers using microarray data. Support vector machine recursive feature elimination (SVM-RFE) is a well known algorithm for this purpose. In this study, a novel gene selection algorithm is proposed to enhance the SVM-RFE method. The proposed approach is designed to use the combination of SVM-RFE and genetic algorithm (GA). The performance of the proposed model is validated on a binary and a multi-category microarray gene expression datasets. The results show that the proposed gene selection method is able to elevate the performance of SVM-RFE, which extracts much less number of informative genes and achieves highest classification accuracy.

Keywords: cancer classification, gene selection, support vector machine recursive feature elimination (SVM-RFE), genetic algorithm (GA), microarray data

1 Introduction

Recent advances in cancer genomic would provide opportunities for personalized cancer medicine [1]. Cancer is a systemic and complex disease with highly heterogeneity, which remains a key obstacle for accurate diagnosis and treatment of cancers. There exist different pathways between patients with tumours, and it is prone to over-treatment or ineffective treatment of the patients if the same type of treatment is used to treat a certain type of cancer. One typical example is the anti-cancer drug Trastuzumab, which is an antibody that interferes with the human epidermal growth factor (HER2) receptor, and is only effective in patients that HER2 is over-expressed [2]. At the same time, personalized cancer medicine underlines the need for molecular classification of tumours and identification of reliable tumour biomarkers to predict tumour subtypes.

Nowadays high throughput technologies such as microarray technology allow for monitoring of thousands of gene expression values simultaneously, and have been successfully conducted in molecular classification of tumours and identification of tumour biomarkers [3]. However, the sample size of microarray data is typically small (less than 100), and the number of genes is large (generally more than 10,000). The key issue that needs to be addressed is to select a smaller number of informative genes from the thousands of genes measured in microarray, which are subsequently used to accurately classify tumour samples [4, 5].

Support vector machine recursive feature elimination (SVM-RFE) is a well known algorithm for this purpose proposed by Guyon et al. [6]. SVM-RFE algorithm has recently attracted many researchers since it can obtain satisfactory results with microarray gene expression data [5, 7-11]. This paper proposes a novel SVM-RFE based gene selection algorithm by combining a genetic algorithm (GA) with SVM-RFE criteria. The proposed method, which is referred to as SVM-RFE/GA, can be divided into two stages: in the first stage, a ranking list for the original gene set is yielded by SVM-RFE criteria, and top n ranking genes are retained as “candidate gene set”; in the second stage, a genetic algorithm is applied on the candidate gene set, in order to search an optimal minimum gene set. Experimental results demonstrate the feasibility and effectiveness of the proposed method. Section 2 describes SVM-RFE and GA methods, and gives a detailed description of the SVM-RFE/GA model. Section 3 demonstrates the experimental results. Section 4 analyses and discusses the results. Section 5 concludes this work.

2 Method

2.1 SVM-RFE

Support vector machine (SVM) is a superior classification model for sparse classification problems such as microarray gene expression data. Due to the high dimensionality of feature space in microarray data, linear SVM is adopted in this work. For a liner SVM, the margin width is defined as:

$$w = \sum_{i=1}^{n} \alpha_i c_i x_i \ , \tag{1}$$

[*] ***Corresponding author*** e-mail: oboaixil@126.com

$$m\arg in\ width = 2/\|w\|, \quad (2)$$

where n is the number of support vectors.

SVM-RFE is a type of embedded gene selection method [4]. SVM-RFE is a backward elimination procedure, which iteratively removes each feature, which is of the least importance to the SVM classifier. The objective function J in SVM-RFE is:

$$J = (1/2)\|w\|^2. \quad (3)$$

The Optimal Brain Damage (OBD) algorithm [12] approximates the change in J caused by removing each feature by expanding J in Taylor series to second order:

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2. \quad (4)$$

The first order can be neglected at the optimum of J, and the second order becomes

$$\Delta J(i) = (\Delta w_i)^2. \quad (5)$$

The change in weight $\Delta w_i = w_i$ correlates with removing i th feature from the classifier, so $(w_i)^2$ is used as the ranking criterion in SVM-RFE. The feature with the smallest $(w_i)^2$ is eliminated since it has the smallest effect on classifier.

The detail of SVM-RFE algorithm is described as follows:

```
Inputs: initial gene set $I = \{1; 2; ...n\}$, ranked gene set $O = \{\ \}$.
While ($I$ is not null)
Train the linear SVM classifier
Calculate the ranking criteria $r_i = (w_i)^2$ for all genes in $I$.
Choose the gene with the smallest ranking score: $g = \arg\min\{r_i\}$
Update ranked gene set $O$ and $I$: $O = O \cup g$, $I = I - g$
End While
Output: ranked gene set $O$
```

2.2 GENETIC ALGORITHM

Genetic algorithms (GA) [13-15] is a type of wrapper gene selection method, which is based on the principle of natural selection and genetics. GA is a globally adaptive probabilistic search algorithm, drawing on the biological mechanisms of fittest evolution and natural selection, and the genetic mechanisms of recombination and mutation. GA starts from an initial solution of randomly generated population, and the population contains a certain number of encoded individuals. Based on the principle of survival of the fittest, the evolution of each generation would produce more and better approximate solutions. In each generation, each individual is evaluated by the fitness function in the solution domain. The more fit individuals are retained and then modified with genetic operators of crossover and mutation, producing a new population representative of the new solution sets. This process loop is executed until a predetermined termination condition has been reached.

The main components of our GA are described as follows.

Representation of individual. Each individual is encoded by a N -bit binary vector, where N is the size of genetic space. The bit "1" represents a selected gene, and the bit "0" means the opposite.

Fitness Function. Each individual is evaluated by a support vector machines (SVM) classifier, i.e., SMO classifier in WEKA [16].GA is designed to minimize the classification error rate.

Genetic Operators. The genetic operations are performed by Roulette wheel selection, single-point crossover, and bit flip mutation.

2.3 THE SVM-RFE/GA MODEL

Among the thousands of genes detected by microarray technology, there exist four categories of genes for cancer classification [10]: (1) informative genes, which are important for cancer classification and may play a significant role in tumour development; (2) redundant genes, which may be related with cancer and function similarly to informative genes but they are not so significant for cancer classification; (3) irrelevant genes, which have no influence on cancer classification and are irrelevant to cancer; and (4) noisy genes, which have negative effects and their existence may decrease cancer classification performance. The gene selection methods are developed to obtain the first class while removing the next three classes of genes.

SVM-RFE eliminates "worst" gene at each step, generating a ranking for the genes based on their "importance" to the classifier. SVM-RFE has achieved an outstanding performance in cancer classification. However, it ignores the interaction between the genes. A two-stage strategy is proposed to overcome this deficiency. In the first stage, SVM-RFE is applied on the initial gene sets to generate ranking for the genes, and top n ranking genes are kept as "candidate gene set". The first stage is considered as a prefiltering process, which is designed to remove redundant, irrelevant and noisy genes while retaining informative genes. In the second stage, since the genes in the candidate gene set may highly correlate with each other, a genetic algorithm is utilized to search an optimal minimum gene set in the solution space.

Based on the spirit of Structural Risk Minimization [6], nested gene subsets are defined by the ranking algorithm, and it is possible to select best gene subset by changing the parameter of n: the number of genes. In more detail, the parameter n is varied to select top n ranking genes, generating m incrementally nested gene subsets: $GS_1 \subset GS_2 \subset ...GS_m$. A genetic algorithm is further applied to search the optimal minimum gene subset from the given input space.

3 Experimental Results

3.1 DATA SET

The performance of the SVM-RFE/GA model is validated on both binary and multi-category microarray gene expression datasets. Table 1 summarizes the number of classes, the number of genes, the number of samples and the reference in each dataset.

TABLE 1 The two-class and multi-class gene expression datasets

Dataset	Platform	No. of Classes	No. of Genes	No. of Samples	Reference
Prostate	Affy U95Av2	2	12600	102	[17]
NCI60	Affy Hu6800	9	7129	60	[18]

The prostate dataset is a two-class gene expression dataset, which contains 52 tumor and 50 normal of prostate cancer samples, and it can be obtained from (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi). The NCI60 dataset is a multi-class gene expression dataset, which can be downloaded from (http://www.broadinstitute.org/mpr/NCI60/). The data set contains 60 samples in 9 tumor types, and the two samples of prostate cancer were excluded from this study, since two samples are not enough for the classification issue.

3.2 EXPERIMENTAL PLATFORM

The experiments were conducted on the WEKA [16] (http://www.cs.waikato.ac.nz/ml/weka/) platform. The SMO classifier was used to execute the classification task, and the polynomial kernel function (PolyKernel) was chosen. The penalty parameter C of the classifier was set to 100. The performance of the SMO classifier was evaluated based on 10-fold cross-validation. The parameters of GA were set as follows: crossover probability = 1, mutation probability = 0.02, maximum generations = 50, and population size = 30.

Pre-processing procedure was performed on the experimental data: the housekeeping genes were removed, with 12,533 gene expression values remained in the prostate dataset and 7,071 gene expression values remained in the NCI60 dataset; The gene expression values were standardized to have a mean of 0 and a standard deviation of 1.

3.3 EXPERIMENTAL RESULTS

In the first stage, SVM-RFE algorithm generates ranked gene set, in which genes rank in descending order. Generally, a smaller subset of 50-100 genes is kept as informative genes in previous study [5]. Here a subset of top 100 ranking genes was retained. To test the performance of SVM-RFE algorithm, the number of genes was reduced from 100 to 1, and the gene with the lowest rank score was eliminated at each step. The performance of the classifiers was assessed using 10-fold cross-validation method. The classifier achieved 100% prediction accuracy with initial 100 genes in both datasets. As shown in Figure 1, in both datasets, the prediction accuracy maintains the highest accuracy when the gene number is reduced. In prostate and NCI60 datasets, the classifiers obtained 100% accuracy with minimum number of 9 and 80 genes, respectively. It was observed that the two-class dataset could obtain satisfactory classification results with less number of genes than the multi-category dataset. In NCI60 dataset, the 10-fold cross-validation accuracy did not exceed 90% when the gene number was less than 36. However, in prostate dataset, the classifier obtained 100% accuracy with minimum number of 9 genes.

To combine SVM-RFE with GA, a different number of top n ranking genes were chosen from the SVM-RFE algorithms as the candidate gene set, where n was set to 10, 20, 30, 50. Since the genetic algorithm is a randomly search model, 5 trials were executed on each candidate gene set, and the results were then averaged.

When Top-10 genes were searched from prostate cancer dataset (Table 2), the genetic algorithm was capable of finding smallest size of subset and achieves 100% classification accuracy. The average subset size of 5.4 genes is less than SVM-RFE method while it needs 9 genes to obtain the same accuracy.

When Top-50 genes were searched from NCI60 cancer dataset (Table 3), the genetic algorithm was capable of achieving 100% classification accuracy. The average subset size of 28 genes is much less than SVM-RFE method while it needs 80 genes to obtain the same accuracy.

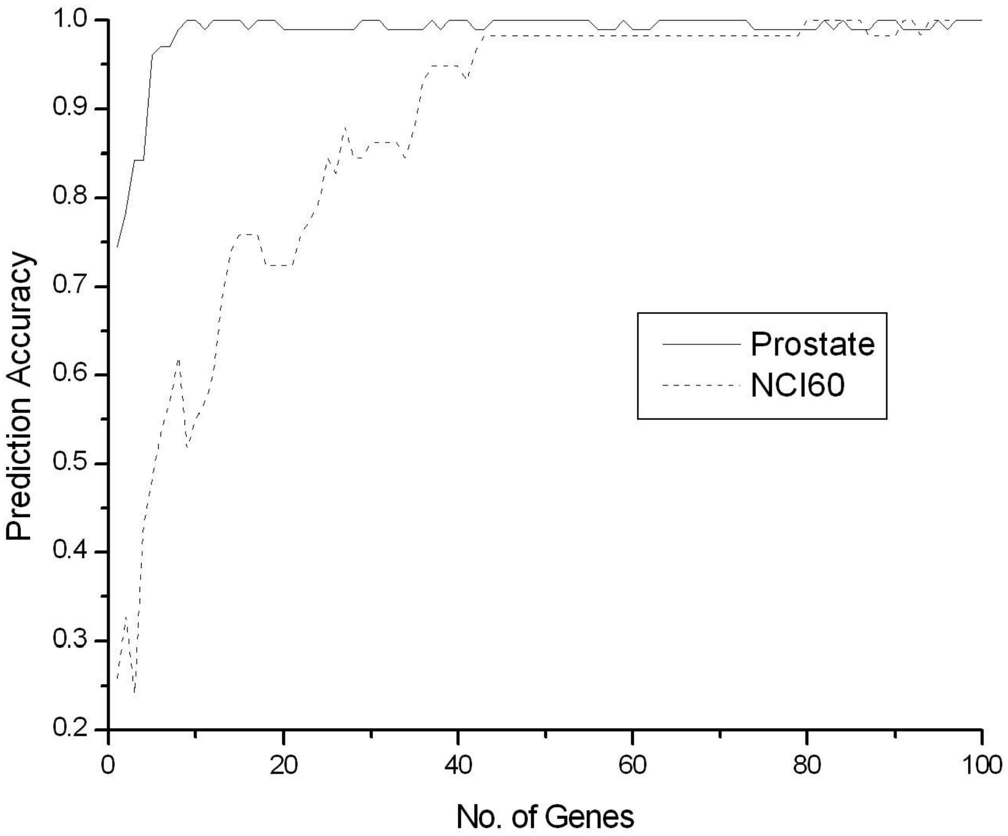

FIGURE 1 The 10-fold cross-validation prediction accuracy in prostate and NCI60 datasets when the number of genes was reduce from 100 to 1

It is observed that the choice of n is a critical problem in GA. When n is too small, the classifier is not able to obtain highest prediction accuracy. On the contrary, when n is too large, GA is possible to be trapped into local optimization, resulting in a larger number of selected genes.

TABLE 2 10-fold accuracy of the SVM-RFE/GA model on prostate cancer data set

Top *n* genes	Average accuracy (%)	Average subset size
10	100	5.4
20	100	7.0
30	100	8.0
50	100	13.2

TABLE 3 10-fold accuracy of the SVM-RFE/GA model on NCI60 cancer data set

Top *n* genes	Average accuracy (%)	Average subset size
10	65.8	6
20	84.6	13.8
30	94.1	20
50	100	28

The gene subset which can achieve highest prediction accuracy with minimum number of genes is defined as the “minimum gene subset”. In prostate cancer dataset, a gene subset selected from Top-10 genes contains minimum number of genes (n=5) while achieving 100% prediction accuracy, and the 5 selected genes are shown in Table 4.

TABLE 4 The selected genes of the minimum subset from prostate cancer dataset

Probe Set ID	Gene Symbol	Gene Title
32786_at	JUNB	jun B proto-oncogene
40282_s_at	CFD	complement factor D (adipsin)
41223_at	COX5A	cytochrome c oxidase subunit Va
41504_s_at	MAF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)
863_g_at	SERPINB5	serpin peptidase inhibitor, clade B (ovalbumin), member 5

In NCI60 cancer dataset, a gene subset selected from Top-50 genes contains minimum number of genes (n=26) while achieving 100% prediction accuracy, and the 26 selected genes are shown in Table 5.

The results of the SVM-RFE/GA model were compared with some other algorithms in two aspects: the prediction accuracy and number of selected genes. In prostate cancer dataset (Table 6), only the SVM-RFE/GA model and SVM-RFE algorithm can achieve 100% prediction accuracy, but SVM-RFE/GA algorithm selects less number of genes.

The performance of SVM-RFE/GA model is more prominent in NCI60 cancer dataset (Table 7), the SVM-RFE/GA algorithm is capable of achieving 100% prediction accuracy, using much less number of genes (n=26) than SVM-RFE algorithm (n=80).

TABLE 5 The selected genes of the minimum subset from NCI60 cancer data set

Probe Set ID	Gene Symbol	Gene Title
AF005775_at	CFLAR	CASP8 and FADD-like apoptosis regulator
AF006041_at	DAXX	death-domain associated protein
D00017_at	ANXA2	annexin A2
D11327_s_at	PTPN7	protein tyrosine phosphatase, non-receptor type 7
D31888_at	RCOR1	REST corepressor 1
HG1869-HT1904_at	/	/
HG2147-HT2217_at	/	/
L38932_at	BECN1	beclin 1, autophagy related
L41349_at	PLCB4	phospholipase C, beta 4
M13929_s_at	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
M37033_at	CD53	CD53 molecule
M59807_at	IL32	interleukin 32
M69181_at	MYH10	myosin, heavy chain 10, non-muscle
M90366_at	ZP2	zona pellucida glycoprotein 2 (sperm receptor)
M93036_at	EPCAM	epithelial cell adhesion molecule
U14577_s_at	MAP1A	microtubule-associated protein 1A
U49250_at	TBR1	T-box, brain, 1
U65785_at	HYOU1	hypoxia up-regulated 1
U95090_at	PRODH2	proline dehydrogenase (oxidase) 2
X03656_rna1_at	CSF3	colony stimulating factor 3 (granulocyte)
X12492_at	NFIC	nuclear factor I/C (CCAAT-binding transcription factor)
X52947_at	GJA1	gap junction protein, alpha 1, 43kDa
X75315_at	RBM38	RNA binding motif protein 38
X91247_at	TXNRD1	thioredoxin reductase 1
Y09305_at	DYRK4	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 4
Z18859_rna1_at	GNAT2	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 2

TABLE 6 Results comparison of the SVM-RFE/GA model with some other algorithms in prostate dataset

Method	Prediction accuracy (%)	No. of Genes	Reference
SVM-RFE/GA	100	5	This study
SVM-RFE	100	9	[6]
SVM-RFE With MRMR	98.29	10	[11]
TSP	95.10	2	[19]
K-TSP	91.18	2	[19]

TABLE 7 Results comparison of the SVM-RFE/GA model with some other algorithms in NCI60 dataset

Method	Prediction accuracy (%)	No. of Genes	Reference
SVM-RFE/GA	100	26	This study
SVM-RFE	100	80	[6]
GA/SVM	87.93	27	[20]
GA/MLHD	85.37	13	[21]

4 Discussions

Gene selection has been a key research issue in microarray data analysis. Gene selection method aims to eliminate noisy, irrelevant and redundant genes, which can not only reduce the computational burden of the classifier, but also improve the classification accuracy of the classifier. In another aspect, the selected informative gene set, which contains least amount of genes, is much easier to be validated in subsequent molecular biology experiments.

As a type of embedded gene selection algorithm which is well coupled with support vector machine classifier, SVM-RFE achieves satisfactory results in the classification of microarray gene expression data. However, it ignores complementary relationship between genes. As a type of wrapper gene selection algorithm, GA selects genes nonlinearly by generating gene subsets randomly, which is efficient in detecting nonlinear relationships among genes. However, GA suffers instability in selecting genes and is prone to be trapped into local optimal solutions as the size of gene subset increases. In this study, a two-stage model is proposed to overcome these limitations.

Gene selection method eliminates noisy, irrelevant and redundant genes, in order to obtain highest classification accuracy with smallest number of genes. The performance of the SVM-RFE/GA model is validated on a binary and a multi-category microarray gene expression datasets. The SVM-RFE/GA model is superior to those previous studies in two aspects: firstly, the SVM-RFE/GA model can obtain highest prediction accuracy, and secondly, the number of selected genes is much less than theirs.

The selected genes in the minimum gene subsets are reported to be associated with cancer development. Out of the 5 selected genes from the prostate dataset (Table 4), JUNB [22] and SERPINB5 [23] were reported to be associated with prostate cancer. JunB is an upstream regulator of p16 and plays a key role in prostate cancer development [22]. Evidence shows that overexpression of SERPINB5 correlates with decreased prostate cancer metastasis [23].

Among the 26 selected genes of the minimum gene subset from NCI60 dataset (Table 5), a fraction of genes are found to have direct associations with cancers. CFLAR plays a role in inhibiting both apoptotic and necroptotic cell death [24]. Death domain-associated protein DAXX promotes ovarian cancer cell proliferation and chemoresistance [25]. Up-regulation of ANXA2 is reported to be associated with poor prognosis in human non-small cell lung cancer [26]. Evidence shows that overexpression of Beclin 1 may inhibit cell growth in colorectal cancer [27]. MYC plays a role in cell cycle progression, apoptosis and cellular transformation [28]. IL32 is expressed in different types of cancer [29]. EPCAM is reportedly to be strongly expressed and associated with breast cancer progression and metastasis

[30]. TXNRD1 is shown to be associated with poor prognosis in breast cancer [31].

5 Conclusions

In summary, this paper presents a model to combine GA with SVM-RFE, in order to enhance the performance of SVM-RFE. The performance of the SVM-RFE/GA model is validated on a binary and a multi-category microarray gene expression datasets. The SVM-RFE/GA model outperforms SVM-RFE algorithm, by taking advantage of both embedded and wrapper approaches. Compared with many previous gene selection algorithms, the SVM-RFE/GA model is capable of finding much smaller sized subsets of informative genes and achieving highest classification accuracy. Many selected genes by SVM-RFE/GA are reportedly associated with cancer, suggesting that SVM-RFE/GA model is an effective tool for molecular cancer classification and identification of cancer biomarkers using microarray data.

Acknowledgments

This study was supported by a grant from the National Natural Science Foundation of China (Grant No. 61373057), and a grant from the Zhejiang Provincial Natural Science Foundation of China (Grant No. Y1110763).

References

[1] Chin L, Andersen J N, Futreal P A (2011) *Nat Med* **17**(3) 297-303

[2] Ong F S, Das K, Wang J, Vakil H, Kuo J Z, Blackwell W L, Lim S W, Goodarzi M O, Bernstein K E, Rotter J I, Grody W W 2012 Personalized *Expert Rev Mol Diagn* **12**(6) 593-602

[3] Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, Coller H, Loh M L, Downing J R, Caligiuri M A, Bloomfield C D, Lander E S 1999) *Science* **286**(5439) 531-7

[4] Inza Y I, Larranaga P 2007 *Bioinformatics* **23**(19) 2507-17

[5] Li X, Peng S, Chen J, Lu B, Zhang H, Lai M 2012 *Biochemical and Biophysical Research Communications* **419**(2) 148-53

[6] Guyon I, Weston J, Barnhill S, Vapnik V 2002 *Machine Learning* **46**(1-3) 389-422

[7] Duan K B, Rajapakse J C, Wang H Y, Azuaje F 2005 *IEEE Transactions on Nanobioscience* **4**(3) 228-34

[8] Zhang X G, Lu X, Shi Q, Xu X Q, Leung H C E, Harris L N, Iglehart D J, Miron A, Liu J S, Wong W H 2006 *BMC Bioinformatics* **7** 197 (13 pages) doi:10.1186/1471-2105-7-197

[9] Zhou X, Tuck D P 2007 *Bioinformatics* **23**(9) 1106-14

[10] Tang Y C, Zhang Y Q, Huang Z 2007 *IEEE-ACM Transactions on Computational Biology and Bioinformatics* **4**(3) 365-81

[11] Mundra P A, Rajapakse J C (2010) *IEEE Transactions on Nanobioscience* **9**(1) 31-7

[12] Le Cun Y, Denker J, Solla S, Touretzky D S 1990 Optimal brain damage *Advances in Neural Information Processing Systems* Morgan Kaufmann 598-605

[13] Tan F, Fu X, Zhang Y, Bourgeois A 2008 *Soft Computing* **12**(2) 111-20

[14] Nicoletta D, Barbara P 2009 An evolutionary method for combining different feature selection criteria in microarray data classification *Journal of Artificial Evolution and applications* **2009** 1-10

[15] Cannas L, Dessi N, Pes B 2011 A Hybrid Model to Favor the Selection of High Quality Features in High Dimensional Domains *Intelligent Data Engineering and Automated Learning - IDEAL 2011* Berlin Heidelberg: Springer 228-35

[16] Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H W 2009 *SIGKDD Explor Newsl* **11**(1) 10-8

[17] Singh D, Febbo P G, Ross K, Jackson D G, Manola J, Ladd C, Tamayo P, Renshaw A A, D'Amico A V, Richie J P, Lander E S, Loda M, Kantoff P W, Golub T R, Sellers W R 2002 *Cancer Cell* **1**(2) 203-9

[18] Staunton J E, Slonim D K, Coller H A, Tamayo P, Angelo M J, Park J, Scherf U, Lee J K, Reinhold W O, Weinstein J N, Mesirov J P, Lander E S, Golub T R 2001 *Proc Natl Acad Sci USA* **98**(19) 10787-92

[19] Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D 2005 *Bioinformatics* **21**(20) 3896-904.

[20] Peng S H, Xu Q H, Ling X B, Peng X N, Du W, Chen L B 2003 *Febs Letters* **555**(2) 358-62

[21] Ooi C H, Tan P 2003 *Bioinformatics* **19**(1) 37-44

[22] Konishi N, Shimada K, Nakamura M, Ishida E, Ota I, Tanaka N, Fujimoto K 2008 *Clin Cancer Res* **14**(14) 4408-16

[23] Luo J L, Tan W, Ricono J M, Korchynskyi O, Zhang M, Gonias S L, Cheresh D A, Karin M 2007 *Nature* **446**(7136) 690-4

[24] Silke J, Strasser A 2013 *Sci Signal* **6**(258) pe2

[25] Pan W W, Zhou J J, Liu X M, Xu Y, Guo L J, Yu C, Shi Q H, Fan H Y 2013 *J Biol Chem* **288**(19) 13620-30

[26] Jia J W, Li K L, Wu J X, Guo S L 2013*Tumour Biol* **34**(3) 1767-71

[27] Chen Z, Li Y, Zhang C, Yi H, Wu C, Wang J, Liu Y, Tan J, Wen J 2013 *Dig Dis Sci* **58**(10) 2887-94

[28] Nair R, Roden D L, Teo W S, McFarland A, Junankar S, Ye S, Nguyen A, Yang J, Nikolic I, Hui M, Morey A, Shah J, Pfefferle AD, Usary J, Selinger C, Baker L A, Armstrong N, Cowley M J, Naylor M J, Ormandy C J, Lakhani S R, Herschkowitz J I, Perou C M, Kaplan W, O'Toole S A, Swarbrick A 2013 *Oncogene* (in print)

[29] Guenin S, Mouallif M, Hubert P, Jacobs N, Krusy N, Duray A, Ennaji M M, Saussez S, Delvenne P (2013) *Molecular Carcinogenesis* n/a-n/a

[30] Martowicz A, Rainer J, Lelong J, Spizzo G, Gastl G, Untergasser G 2013 *Mol Cancer* **12** 56

[31] Cadenas C, Franckenstein D, Schmidt M, Gehrmann M, Hermes M, Geppert B, Schormann W, Maccoux LJ, Schug M, Schumann A, Wilhelm C, Freis E, Ickstadt K, Rahnenfuhrer J, Baumbach JI, Sickmann A, Hengstler J G 2010 *Breast Cancer Res* **12**(3) R44

Author

Xiaobo Li

Current position, grades: full-time Assoc. Professor, Ph.D.
University studies: B.Sc. in Microelectronics (1990) from Nankai University (China), Master of Engineering (Research) (2004) from The University of Sydney (Australia) and Ph.D. in Pathology and Pathophysiology (2012) from Zhejiang University (China)
Research interests: different aspects of bioinformatics, machine learning and data mining