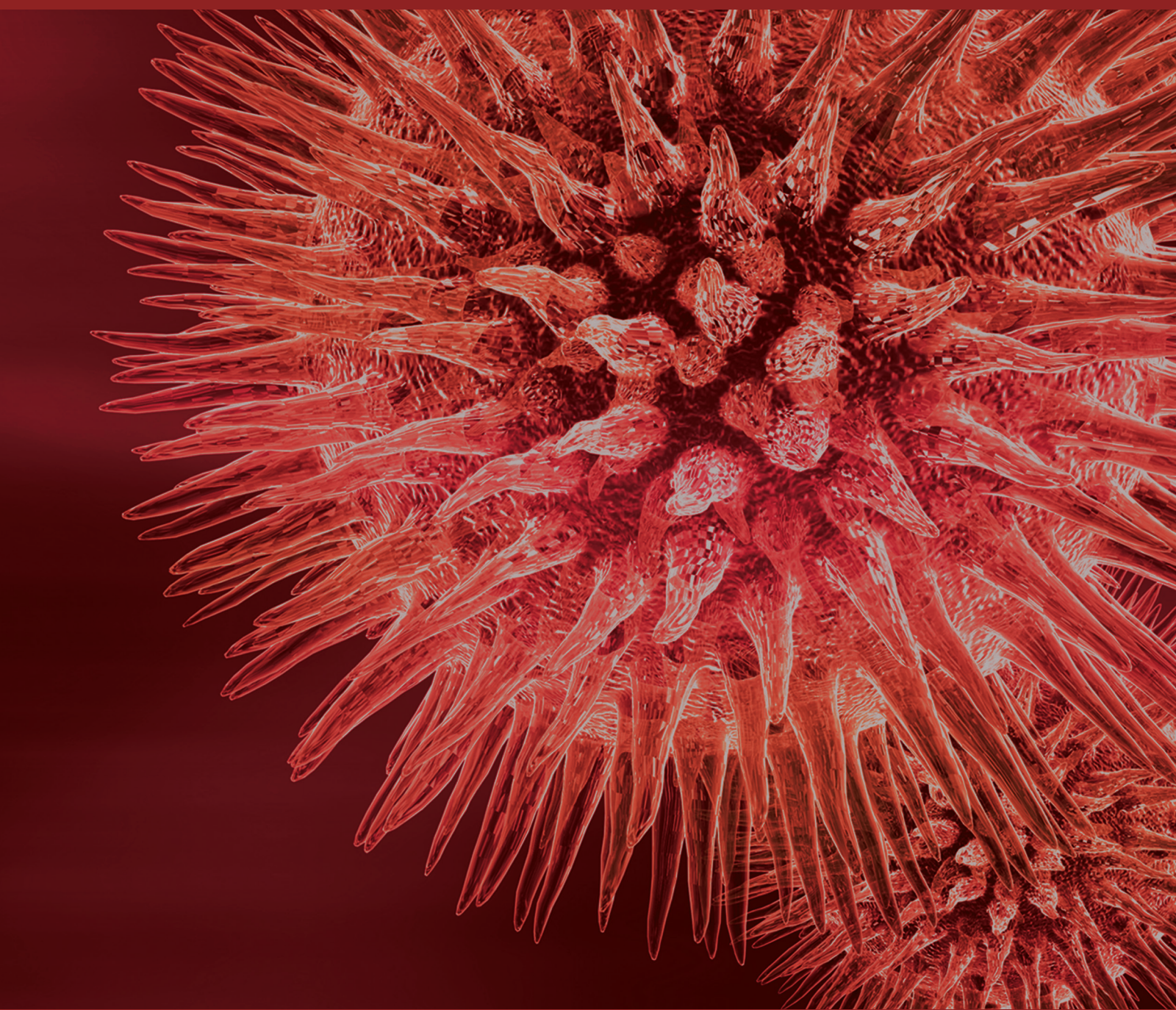
BioMed Research International

Analysis and Modeling for Big Data in Cancer Research

Lead Guest Editor: Bing Niu
Guest Editors: Peter B. Harrington, Guozheng Li, Jianxin Li, and Simon Poon



Analysis and Modeling for Big Data in Cancer Research

BioMed Research International

Analysis and Modeling for Big Data in Cancer Research

Lead Guest Editor: Bing Niu
Guest Editors: Peter B. Harrington, Guozheng Li, Jianxin Li, and Simon Poon

Contents

Hindawi
BioMed Research International
Volume 2017, Article ID 1972097, 2 pages
https://doi.org/10.1155/2017/1972097

Editorial

Analysis and Modeling for Big Data in Cancer Research

Bing Niu,[1] Peter B. Harrington,[2] Guozheng Li,[3] Jianxin Li,[4] and Simon Poon[5]

[1]*College of Life Science, Shanghai University, Shanghai, China*
[2]*Center for Intelligent Chemical Instrumentation, Department of Chemistry and Biochemistry, Ohio University, Athens, OH, USA*
[3]*China Academy of Chinese Medical Sciences, Beijing, China*
[4]*Big Data Research Group, School of Computer Science & Software Engineering, The University of Western Australia (Go8), Perth, WA, Australia*
[5]*School of Information Technologies, University of Sydney, Sydney, NSW, Australia*

Correspondence should be addressed to Bing Niu; phycocy@163.com

Received 12 April 2017; Accepted 12 April 2017; Published 12 June 2017

Cancer is a major disease which has become the biggest threat to human health due to its difficult early detection, diagnosis, and treatment. According to the survey of the World Health Organization in 2012, there were four million new cancer cases and 8.2 million cancer-related deaths worldwide. The history of treatment of tumors covered traditional herbal medicines, surgical anatomy, antitumor chemotherapy/radiotherapy, and new targeted drug therapy and immunotherapy. In the past few decades, with the rapid development of high-throughput technologies such as microarrays and next-generation sequencing (NGS), increasing in-depth studies of tumor biology were spurred at the genetic and genomic level, leading to better targeted and personalized healthcare solutions for cancer patients. The successful implementation of the human genome project has made people realize that genetic, environmental, and lifestyle factors should be combined together to study cancer due to its complexity. For example, some malignant tumors have been proven to be related to the mutations of a drive gene by using specific monoclonal antibodies and small molecule compounds to block or suppress the relevant molecular targets that can inhibit tumor growth and metastasis or induce apoptosis; the survival time of patients has been significantly extended.

The increasing availability and growth rate of "Big Data" derived from various omics open a new window to improve clinical diagnoses or therapeutics of cancer, but there are many challenges in efficient analysis and interpretation of such big and complex data. For instance, how to manage, extract, analyze, integrate, visualize, and communicate the hidden information from the myriad of data representations of cancer evolved into one of the greatest challenges in next-generation biomedicine. Thus, there is a need to fundamentally address all the above-mentioned issues in Big Data in cancer healthcare.

There are six interesting research papers in this special issue covering machine learning methods on feature selection of gene expression profile, cancer prediction, potential new drug design, and QSAR study of anticancer drugs.

Gene expression profiles provide a new insight into cancer diagnosis at a molecular level which paved the way towards personalized medicine. Gene expression data usually contains a large number of genes, but a small number of samples. Feature selection for gene expression data aims at finding a set of genes that best discriminate biological samples of different types. Q. Su et al. proposed a gene subset selection algorithm based on the Kolmogorov-Smirnov (K-S) test and correlation-based feature selection (CFS) principles to address the challenging problem of selecting distinguished genes from cancer gene expression datasets. The authors compared the K-S test plus CSF with K-S test alone, CFS alone, ReliefF, and mRMR feature selection in 5 cancer gene expression datasets, which adopted support vector machines (SVM) as the classification tool and used the criteria of accuracy to evaluate the performance of the classifiers on the

selected gene subsets. The results show that this combination algorithm is more efficient.

L. Yang et al. presented a gene subset selection algorithm RS_SVM based on aggregating SVMs trained on eight random subspaces gene expression profiles. The results show that RS_SVM outperforms single SVM, KNN, CART, Bagging, AdaBoost, and 16 state-of-the-art methods in literatures. L. Yang et al.'s study provides a potential tool for the problems of high dimension and small sample problem in gene expression data which could lead to overfitting and huge computing pressure. The authors also proposed that RS_SVM is not suitable for heterogeneous data as they failed to apply RS_SVM with PCA on two gene expression profiles.

Glioma is the most common and most aggressive malignant brain tumor in humans that affects nonneural glial cells in the central nervous system. The knowledge of glioblastoma at the molecular and structural level will greatly improve the treatment of glioma in the clinic. H. Long et al. made a PPI network of key DEGs to study the significant functions associated with the occurrence and development of glioma combined with enriched GO and KEGG data. Pathways in cancer, MAPK signaling pathway, focal adhesion, and calcium signaling pathway were regarded to be related to the occurrence of glioma. In addition, some key genes such as MMP9, CD44, CDC42, COL1A1, COL1A2, CAMK2A, and CAMK2B were also proposed, which might be target genes for diagnosing glioblastoma.

EGFR is considered to be an anticancer target as it has been found in some solid tumors, such as glioma, lung cancer, ovarian cancer, breast cancer, and other cancers. Several efforts have been made to develop EGFR inhibitors for the treatment of cancer. The low selectivity, high toxicity, and reduced activity promote the design of improved EGFR. M. Zhao et al. introduced the application of 2D and 3D QSAR methods to discriminate EGFR inhibitors and subsequently performed structural docking of the molecules. Overall, this study is modest but nice which contributes to a deeper understanding of the intricacies of drug potency for inhibiting EGFR.

J. Chen et al. developed a novel monocarbonyl curcumin analog which exhibits preferable anticancer effects on laryngeal cancer cells via targeting NF-κB with little toxicity to normal cells. Many traditional Chinese medicine extracts have preferable anticancer effects; however, their toxicities are usually neglected. Meanwhile, this study also reveals that NF-κB is probably a potential target for laryngeal cancer treatment using molecular docking method. Therapeutics based on targeting NF-κB may be effective approaches for laryngeal cancer treatment in the future. However, the results of this study all come from in vitro trials; further tests of this curcumin analog in vivo need to be performed.

M.C. Ng et al. built a bioactivity model for complex mixtures of herb Radix Astragali (RA) extracts based on chemical fingerprinting profiles with Elastic Net Partial Least Square (EN-PLS) algorithm. The prediction platform they obtained has the capacity to identify potential key bioactivity-related chemical components of the herb, which is helpful for discovering potential novel drugs, especially for the herbal extracts to be used in clinical trials.

Bing Niu
Peter B. Harrington
Guozheng Li
Jianxin Li
Simon Poon

Hindawi
BioMed Research International
Volume 2017, Article ID 4649191, 11 pages
https://doi.org/10.1155/2017/4649191

Research Article

2D-QSAR and 3D-QSAR Analyses for EGFR Inhibitors

Manman Zhao,[1] Lin Wang,[2] Linfeng Zheng,[3] Mengying Zhang,[1] Chun Qiu,[2] Yuhui Zhang,[4] Dongshu Du,[1,5] and Bing Niu[1]

[1]*Shanghai Key Laboratory of Bio-Energy Crops, College of Life Science and Shanghai University High Performance Computing Center, Shanghai University, Shanghai 200444, China*
[2]*Department of Oncology, Hainan General Hospital, Haikou, Hainan 570311, China*
[3]*Department of Radiology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China*
[4]*Changhai Hospital, Second Military Medical University, Shanghai 200433, China*
[5]*Department of Life Science, Heze University, Heze, Shandong 274500, China*

Correspondence should be addressed to Yuhui Zhang; gong_chang2008@126.com, Dongshu Du; dsdu@shu.edu.cn, and Bing Niu; phycocy@163.com

Received 6 January 2017; Revised 19 February 2017; Accepted 12 March 2017; Published 29 May 2017

Academic Editor: Vladimir Bajic

Epidermal growth factor receptor (EGFR) is an important target for cancer therapy. In this study, EGFR inhibitors were investigated to build a two-dimensional quantitative structure-activity relationship (2D-QSAR) model and a three-dimensional quantitative structure-activity relationship (3D-QSAR) model. In the 2D-QSAR model, the support vector machine (SVM) classifier combined with the feature selection method was applied to predict whether a compound was an EGFR inhibitor. As a result, the prediction accuracy of the 2D-QSAR model was 98.99% by using tenfold cross-validation test and 97.67% by using independent set test. Then, in the 3D-QSAR model, the model with $q^2 = 0.565$ (cross-validated correlation coefficient) and $r^2 = 0.888$ (non-cross-validated correlation coefficient) was built to predict the activity of EGFR inhibitors. The mean absolute error (MAE) of the training set and test set was 0.308 log units and 0.526 log units, respectively. In addition, molecular docking was also employed to investigate the interaction between EGFR inhibitors and EGFR.

1. Introduction

Epidermal growth factor receptor (EGFR), a transmembrane glycoprotein, is classified to the prototype of receptor tyrosine kinases (TKs) family that includes EGFR, ErbB-2, ErbB-3, and ErbB-4. EGFR is activated by its cognate ligands via forming a homodimer or heterodimer with other members of the EGFR family, such as epidermal growth factor (EGF) and transforming growth factor alpha (TGF-α) [1]. Several signal transduction cascades are initiated when EGFR is activated and then lead to DNA synthesis and cell proliferation [2, 3]. While EGFR is amplified or mutated, DNA synthesis and cell proliferation will be abnormal and lead to cancer. Currently, the amplification or mutation of EGFR has been found in human solid tumors, such as glioma, lung cancer, ovarian cancer, and breast cancer. Hence, EGFR is also considered to be a potential anticancer target in this disease [4–8]. Many EGFR inhibitors have been developed and approved by the FDA, such as lapatinib, which has been applied for the treatment of breast cancer [9]. Moreover, other EGFR inhibitors like temozolomide, lomustine, erlotinib, and gefitinib, are approved by the FDA for the treatment of glioma [10, 11]. However, the existing EGFR inhibitors are beyond people's expectation due to selectivity, toxicity, and side effect. Hence, it is necessary to design and synthesize new potential EGFR inhibitors.

Quantitative structure-activity relationship (QSAR) was a valuable tool for many different applications, including drug discovery, predictive toxicology, and risk assessment [12–14]. The applicability domain of QSAR models, defined by the Organization for Economic Co-operation and Development (OECD) according to Principle 3, includes the physicochemical, the structural, and the biological domain [15–17]. Initially, two-dimensional quantitative structure-activity relationship (2D-QSAR) was widely explored and used in medicinal chemistry study. However, some limitations spurred

the appearance of three-dimensional quantitative structure-activity relationship (3D-QSAR). In the 3D-QSAR study, the correlation between 3D steric and electrostatic fields and biologically activity draws attention. For the molecular field study, CoMFA was widely used preliminarily. However, the time-consuming limit stimulates the advent of TopCoMFA. TopCoMFA overcomes the weakness and uses an objective method to fragment and align the molecules. In addition, the fragmentation process is automated except for some specific bonds that should be cleaved manually. Of course, TopCoMFA and CoMFA also have similarity that they both share QSAR PLS analysis. The details about TopCoMFA and CoMFA are in [18].

Drug development is a long process, and it requires a vast amount of material and financial resources. QSAR and molecular docking technology have been extensively employed in drug virtual screening and potential molecular targets prediction, which may shorten the cycle of the drug development [19–22]. In this work, 2D-QSAR model was employed to determine EGFR inhibitor, and the 3D-QSAR model was used to predict the activity. Finally, molecular docking was applied to investigate the binding sites.

2. Materials and Methods

2.1. CfsSubsetEval Method and Greedy Stepwise Algorithm. A data set containing n vectors has 2^n possible combinations of features for the subset. A useful subset which can correctly predict other compounds is one of 2^n combinations. The best way to find an optimal subset is to try all the possible feature combinations. However, this strategy is difficult to carry out due to the huge computation. In this study, the CfsSubsetEval (CFS) search method combined with Greedy Stepwise (GS) algorithm was employed to search the optimal feature subset. The main idea of the GS algorithms is to make the best choice when selecting good features. The CFS method was used to evaluate the attribute. Thus, the CFS method, combined with the GS algorithm, was employed to select the optimal subset from these 2^n combinations. Additional details about the CFS method and the GS algorithm could be found in [23–25].

2.2. SVM. Support vector machine (SVM), a supervised learning algorithm, is usually used for pattern recognition classification [26]. SVM was employed for the classification and sensitivity analysis in our study due to its high performance in many studies [25, 27, 28].

2.3. Topomer CoMFA. Topomer CoMFA, possessing both the topomer technique and CoMFA technology, can overcome the alignment problem of CoMFA [18, 29]. Partial least squares (PLS) regression is employed to build the topomer CoMFA model, and the leave-one-out (LOO) cross-validation is used to evaluate the model. Additional details about the topomer CoMFA can be found in [29–31].

2.4. Data Preparation. 100 inhibitors derived from the literature and 185 noninhibitors downloaded from the DUD database (http://dud.docking.org) were collected [32–41]. For 2D-QSAR study, the data set containing inhibitors and noninhibitors was randomly divided into three training sets which accounted for 75%, 70%, and 50% of the whole data set, respectively (see Supplementary Material 1, available online at https://doi.org/10.1155/2017/4649191). For 3D-QSAR study, the 100 inhibitors were randomly divided into a training set (77 molecules) and an independent test set (23 molecules).

2.5. Molecular Descriptor Calculation. Molecular descriptor can reflect physicochemical and geometric properties of the compounds. In this study, forty-five molecular descriptors calculated by the ChemOffice were applied to represent compounds [42]. First, three-dimensional structures of the molecules were optimized by MM+ force field with the Polak-Ribiere algorithm until the root-mean-square gradient became less than 0.1 Kcal/mol. Then, quantum chemical parameters were obtained for the most stable conformation of each molecule by using PM3 semiempirical molecular orbital method at the restricted Hartree-Fock level with no configuration interaction.

2.6. Validation Methods for Prediction Results. In this study, tenfold cross-validation test and independent set test were applied to evaluate the prediction ability of the 2D-QSAR model. For the tenfold cross-validation test, the data set was divided into ten subsets. Nine subsets were used as the training set and the left subset was predicted. In turn, each subset was omitted in order to be predicted, and the correct rate was obtained from each trial. The average of the correct rate from ten trials was used to estimate the accuracy of the algorithm [43–45].

2.7. Prediction Measurement. Sensitivity (SN), specificity (SP), overall accuracy (ACC), and Matthew's correlation coefficient (MCC) were employed to evaluate the 2D prediction model. The SN, SP, ACC, and MCC can be represented as

$$\begin{aligned} \mathrm{SN} &= \frac{\mathrm{TP}}{[\mathrm{TP}+\mathrm{FN}]}, \\ \mathrm{SP} &= \frac{\mathrm{TN}}{[\mathrm{TN}+\mathrm{FP}]}, \\ \mathrm{ACC} &= \frac{[\mathrm{TP}+\mathrm{TN}]}{[\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}]}, \\ \mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})}}. \end{aligned} \tag{1}$$

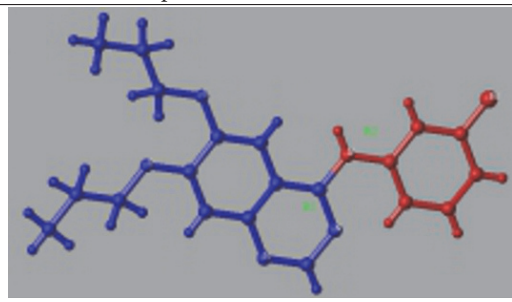
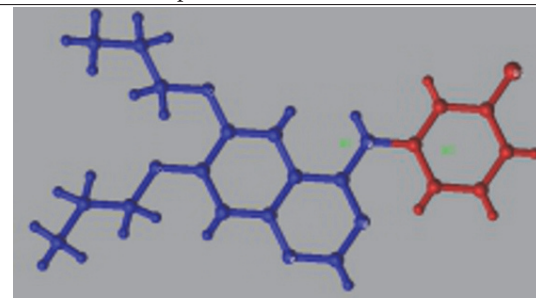
TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively.

In the topomer CoMFA model, q^2, r^2, and MAE were applied to evaluate the model [46]. The cut-off value of q^2 is 0.5. The MAE of the test set was less than 0.1 × training set range and MAE + 3 × σ according to the MAE based criteria. The optimized model was determined by the highest q^2, and the validity of the model depends on r^2 value [47].

Table 1: The results of prediction accuracy for different data sets containing 9 molecular descriptors using SVM classifier. DS and EP present data set and evaluation parameters, respectively.

EP	DS			
	Train set (75%)	Train set (70%)	Train set (50%)	Test set (30%)
SN (%)	97.22	98.55	91.94	96.77
SP (%)	98.59	99.23	90.67	98.18
ACC (%)	98.13	98.99	91.24	97.67
MCC	0.958	0.978	0.824	0.950

Table 2: Results from two topomer CoMFA model studies.

Dataset	Topomer CoMFA model 1	Topomer CoMFA model 2
Cutting model		
q^2	0.483	0.565
r^2	0.773	0.888

2.8. Steric and Electrostatic Field Analysis. Topomer CoMFA analysis is an effective approach which has been applied in drug design for HIV, central nervous system diseases, and other tumors [48–50]. In the topomer CoMFA model, there are two different ways to calculate the molecular field. One way is to reduce the field contributions of fragmenting atoms; the other way is to calculate the steric and electrostatic fields on a regularly spaced grid. For detailed information, see [51]. Topomer CoMFA analysis is used to calculate the steric field and electrostatic fields of R1 and R2 groups. Steric and electrostatic field analysis may help design novel EGFR drugs.

2.9. Molecular Docking. SYBYL X-2.0 was used for molecular docking based on its Surflex-Dock module [52]. The crystal structure of EGFR with the resolution of 2.6 Å was downloaded from the Protein Data Bank (PDB ID: 1M17) [53]. Protein was prepared with protein structure preparation module of the SYBYL X-2.0. All the water molecules and ligands were deleted, and hydrogen atoms were added to the crystal structure. In addition, positive and negative charges were added to N-terminal and C-terminal regions of the EGFR which became NH^{3+} and COO^{-}. EGFR inhibitors were minimized at physiological pH 7.0 with hydrogen atoms and charge by using Powell energy gradient method and the Gasteiger-Huckel system.

3. Results

3.1. Feature Selection and the 2D-QSAR Prediction Model. A feature subset containing nine molecular descriptors (DPLL, H, HF, HOMO, MR, Pc, TIndx, VP, and WIndx) was obtained based on CFS combined with GS algorithms. Sensitivity analysis was applied to these nine descriptors to evaluate how they affected the activity of EGFR inhibitors (see Figure 1).

Based on the optimal features subset, the SVM classifier method was used to build the 2D-QSAR prediction model. As a result, the prediction accuracy of these models whose data set accounted for 75%, 70%, and 50% of the whole data set was 98.13%, 98.99%, and 91.24%, respectively, by tenfold cross-validation test. The sensitivity, specificity, and overall accuracy of these three models were more than 90%, which indicated that changing the size of the training set had a little impact on the quality of the 2D-SAR models (see Table 1). The model built via the data set accounting for 70% of the whole data set was chosen finally due to its higher prediction accuracy and smaller size. Although the result of the tenfold cross-validation test was well, it was not good enough for evaluating the classifier as the SVM classifier might be overfitted. To validate the reliability of the classifier, an independent test set was further employed in this study. As a result, the prediction accuracy of the independent set test was 97.67%.

3.2. 3D-QSAR Prediction Model. The training set was employed to build the topomer CoMFA model by fragmenting EGFR inhibitors into R1 and R2 groups. Two topomer CoMFA models were generated by two cutting ways. The topomer CoMFA model 2 with higher q^2 and r^2 values was selected to analyze and predict EGFR inhibitors' activities (see Table 2).

The experimental and predicted activities of the training set and the independent test set were listed in Table 3 and

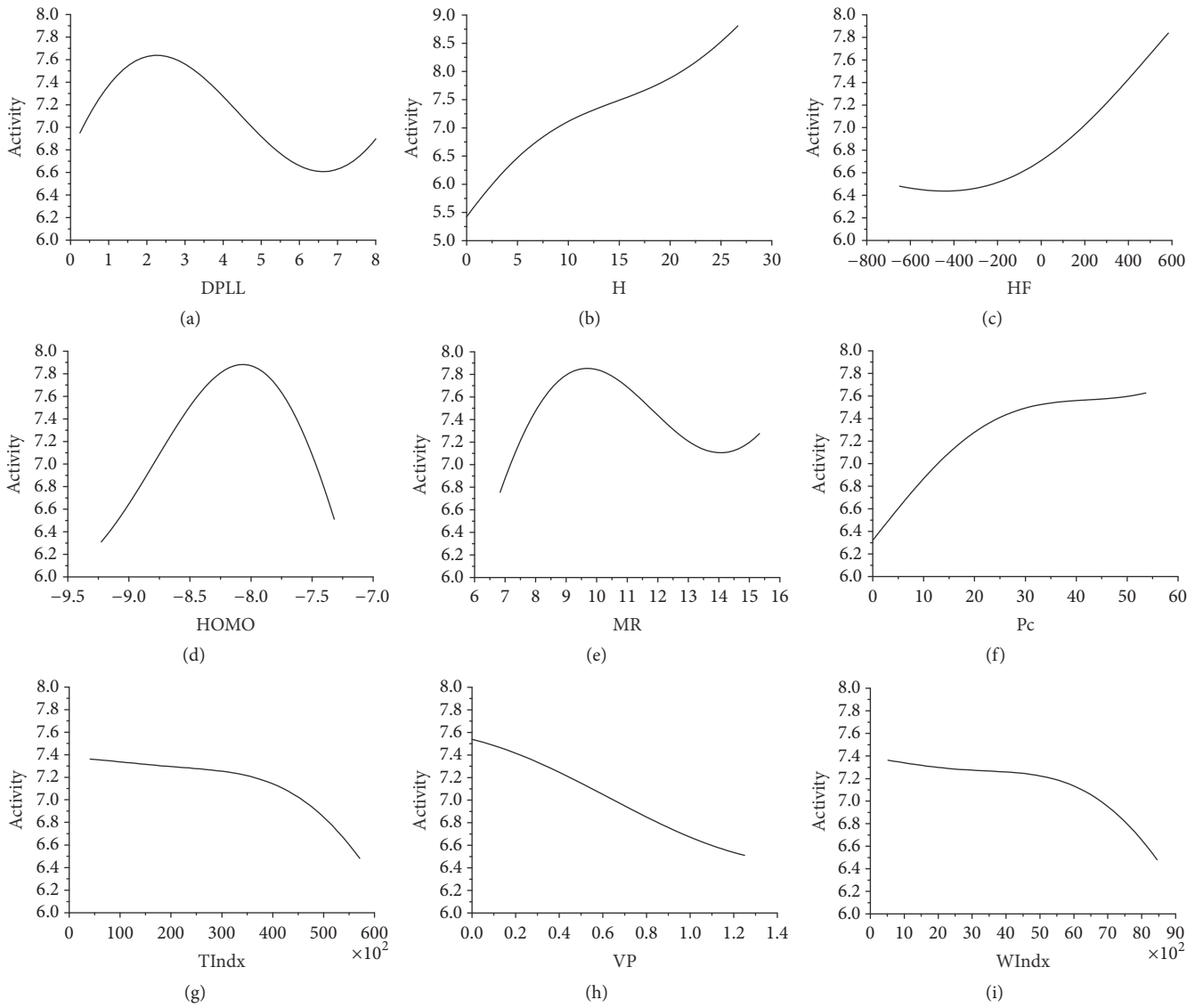


Figure 1: (a) Activity value versus DPLL. (b) Activity value versus H. (c) Activity value versus HF. (d) Activity value versus HOMO. (e) Activity value versus MR. (f) Activity value versus Pc. (g) Activity value versus TIndx. (h) Activity value versus VP. (i) Activity value versus WIndx.

Figure 2. As a result, the MAE and r^2 of the training set were 0.308 and 0.888, respectively. The training set range was 7.32. To estimate the reliability of model 2, the independent set test was used to evaluate the model. The MAE and r^2 of the test set were 0.526 and 0.681, respectively. The MAE of the test set was less than 0.732 (0.1 × training set range) and 1.903 ($\text{MAE}_{(\text{training set})} + 3 \times \sigma$).

Additionally, steric and electrostatic contour maps of R1 and R2 groups were obtained. Compound **33** was selected to study how to redesign EGFR inhibitors due to the highly activity (see Figure 3). From Figure 3, it could be concluded that large volume and positively charged groups were added, which can increase compound activity.

3.3. Molecular Docking. Compounds **27**, **28**, **30**, **31**, **32,** and **33** were used for molecular docking with EGFR. As a result, these compounds have hydrogen bonds at Thr766 and Met769 which were in ATP binding sites (see Figure 4). These compounds interact with EGFR kinase at binding sites and the quinolone ring bound to the hydrophobic pocket of EGFR, instead of the purine ring of ATP.

4. Discussion

4.1. 2D-QSAR Model. Feature selection via removal of some unnecessary features is required for a precise prediction model [25, 54, 55]. A subset containing nine features was obtained to build the 2D-QSAR prediction model. The prediction accuracy of the model was well for the training set and independent test set. This result indicated that the original data contained some redundant features, and feature selection was a helpful step in building a prediction model.

Table 3: Experimental and predicted PIC_{50} for topomer CoMFA model 2.

Compound	Exp	Pre
	Training set	
2	7.64	6.62
4	6.24	6.2
5	6.04	6.45
7	6	6.16
8	8	8.15
10	7.25	7.05
11	6.11	6.62
13	7	6.31
15	6.09	6.06
16	6.26	6.14
17	7.53	8.02
18	9.5	9.06
20	8.39	8.28
22	7.92	8.01
23	8.32	7.59
24	8.15	8.05
25	7.92	8.22
26	7.95	7.78
27	9.16	8.64
29	8.42	8.87
30	8.18	8.3
31	7.82	8.03
32	7.6	7.26
33	9.76	9.6
34	9.01	8.05
36	8.11	7.94
37	7.74	7.43
38	7.35	7.31
40	8.01	8.59
41	8.36	8.46
42	7.45	7.71
43	7.88	7.7
45	6.6	6.36
46	7.39	7.84
47	8	7.5
48	7.04	6.87
50	6.88	6.82
51	6.17	6.08
53	5.74	6.36
54	5.31	5.72
55	6.07	7.21
56	6.92	7.4
57	7.39	6.9
58	7.29	7.14
60	6.9	7.15
61	8.58	8.47
63	6.16	5.85
64	6.02	6.36
65	7.28	6.86
66	6.48	6.54
67	6.58	7
69	7.08	7.57
70	8.82	8.38
71	9.11	8.97
72	9.02	8.97
73	8.42	8.96
75	8.53	9.2
76	8.63	8.35
77	6.42	6.97
78	7.76	7.78
79	8.36	8.34
80	8.63	8.39
81	6.19	6.8
82	8.52	7.97
83	8.05	8.04
85	7.1	7.16
86	7.5	7.57
87	7.26	7.52
88	6.04	6.06
90	4.33	4.35
91	4.66	4.62
92	5	5.52
94	7.19	7.17
95	6.23	5.89
97	4.14	3.98
98	8.05	7.54
99	6.97	6.79
	Test set	
1	6.46	5.58
3	7.57	7.72
6	6.45	6.16
9	7.25	6.44
12	6.24	7.24
14	5.21	5.87
19	9.05	9.14
21	7.07	7.41
28	6.79	7.33
35	7.46	7.23
39	8.5	8.53
44	7.4	8.31
49	5.43	6.4
52	5.27	6.36
59	7.39	7.25
62	8.63	8.41

Table 3: Continued.

Compound	Exp	Pre
68	7.88	7.81
74	9.09	8.98
84	6.72	7.69
89	5.94	5.67
93	7.17	6.68
96	5.01	6.82
100	6.2	6.18

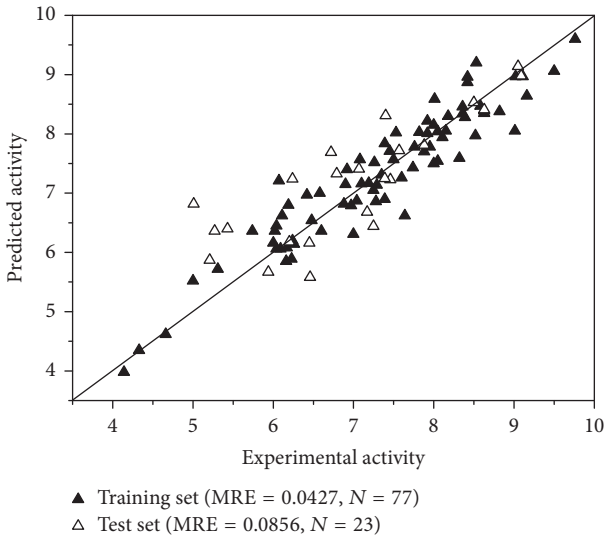
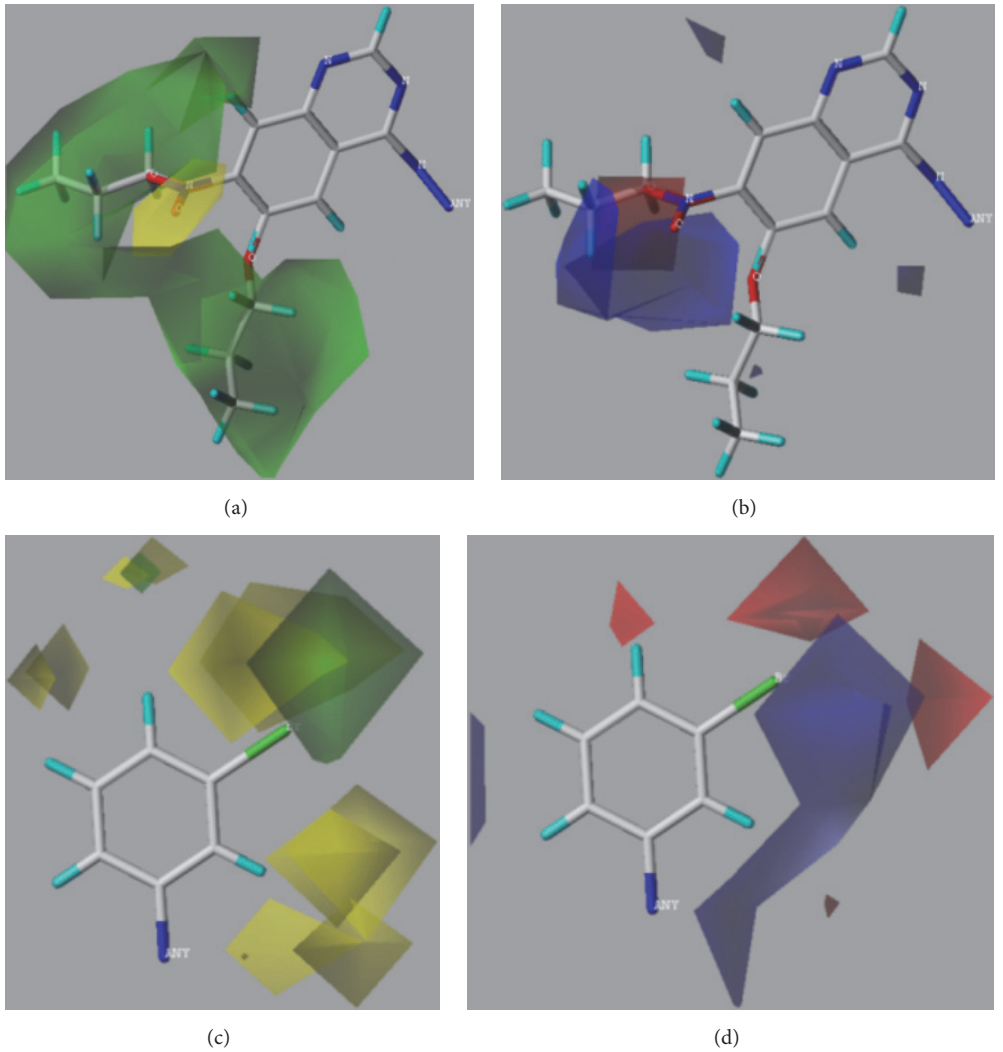


Figure 2: Scatterplot of experimental data versus predicted data from topomer CoMFA model 2.

Although the accuracy of the prediction model with a subset containing nine features (DPLL, H, HF, HOMO, MR, Pc, TIndx, VP, and WIndx) was reliable, it was difficult to analyze the relationship between these descriptors and the activity of EGFR inhibitors as the prediction model is nonlinear. Thus, sensitivity was further applied for this problem [56]. Figure 1(a) shows the relationship between the Dipole length and activity. When the Dipole length is approximately 2 and 6.5, the activity levels are at minimum and maximum, respectively. Figure 1(b) shows the relationship between Henry's law constant and activity. The activity increases along with Henry's law constant from 0 to 30. When Henry's law constant is more than 30, the activity has a rising trend. Figure 1(c) shows the relationship between the Heat of Formation and activity. When the Heat of Formation ranges from −700 to 600, the activity increases. When the Heat of Formation is more than 600, the activity has a rising trend. Figure 1(d) shows the relationship between the HOMO energy and activity. When the HOMO energy ranges from −9.25 to −8.25, the activity increases. When the HOMO energy is approximately −8.25, the activity peaks. When the HOMO energy is greater than −8.25, the activity decreases. When the HOMO energy is more than −7.25, the activity has a decreasing trend. Figure 1(e) shows the relationship between the Molar refractivity and activity. When the Molar refractivity is approximately 10 and 14, the activity levels are at minimum and maximum, respectively. Figure 1(f) shows the relationship between the critical pressure and activity. When the critical pressure ranges from 0 to 60, the activity increases. When the critical pressure is more than 60, the activity has a rising trend. Figure 1(g) shows the relationship between the molecular topological index and activity. When the molecular topological index ranges from 0 to 60,000, the activity decreases. When the molecular topological index is more than 60,000, the activity has a decreasing trend. Figure 1(h) shows the relationship between the Vapor pressure and activity. When the Vapor pressure ranges from 0 to 1.4, the activity decreases. When the Vapor pressure was more than 1.4, the activity had a decreasing trend. Figure 1(i) shows the relationship between the Wiener index and activity. When the Wiener index and activity range from 0 to 9,000, the activity decreases. When the Wiener index is more than 9,000, the activity has a decreasing trend.

4.2. 3D-QSAR Model. Molecules in the topomer CoMFA models can be split into two, three, four, and more groups as needed [51, 57]. In this study, compounds were divided into two groups (R1 and R2). EGFR inhibitors' activity was related to the completeness of the pharmacophore. In topomer CoMFA models, the pharmacophore is related to cutting [44, 48, 58], which plays an important role in the model's predictive performance of the model [58]. In the topomer CoMFA analysis, all molecules of the training set are cut into two fragments. While the fragmentation was complete, the input structures were standardized and the topomers were generated. They all shared the same identical substructure. If the same identical substructure was recognized by the test set, the model's predictive ability was promising.

It could be found that model 2 added an N element in R1 based on model 1, which contributed to the model's predictive ability (see Table 2). Thus, it is speculated that R1 and R2 in model 2 are the same identical substructures. The independent set test was used for evaluating model 2 (see Figure 2). It was observed that the predicted pIC_{50} of some compounds was poor, such as compound **9** and compound **34** (see Table 3). We guess this is because the same identical substructures of the two compounds (see Figure 5) were different from the other compounds. The poor predicted pIC_{50} of compounds may cause high MAE. According to Roy et al.'s report [46], the 3D-QSAR model in our study was reliable as the MAE of the external validation was both less than 0.1 × training set range and MAE (training set) + 3 × σ. It is well known that the presence of systematic error in predictions may easily be identified from the difference in mean error and mean absolute error. It is important to analyze prediction errors of compounds in test set in order to search any possible systematic error. In Roy et al.'s study [59], various metrics, including the number of positive prediction errors (NPE), the number of negative prediction errors (NNP), the absolute value for average of prediction errors (AE), the average of absolute prediction errors (AAE), the mean of positive prediction errors (MPE), and the absolute value for mean of negative prediction errors



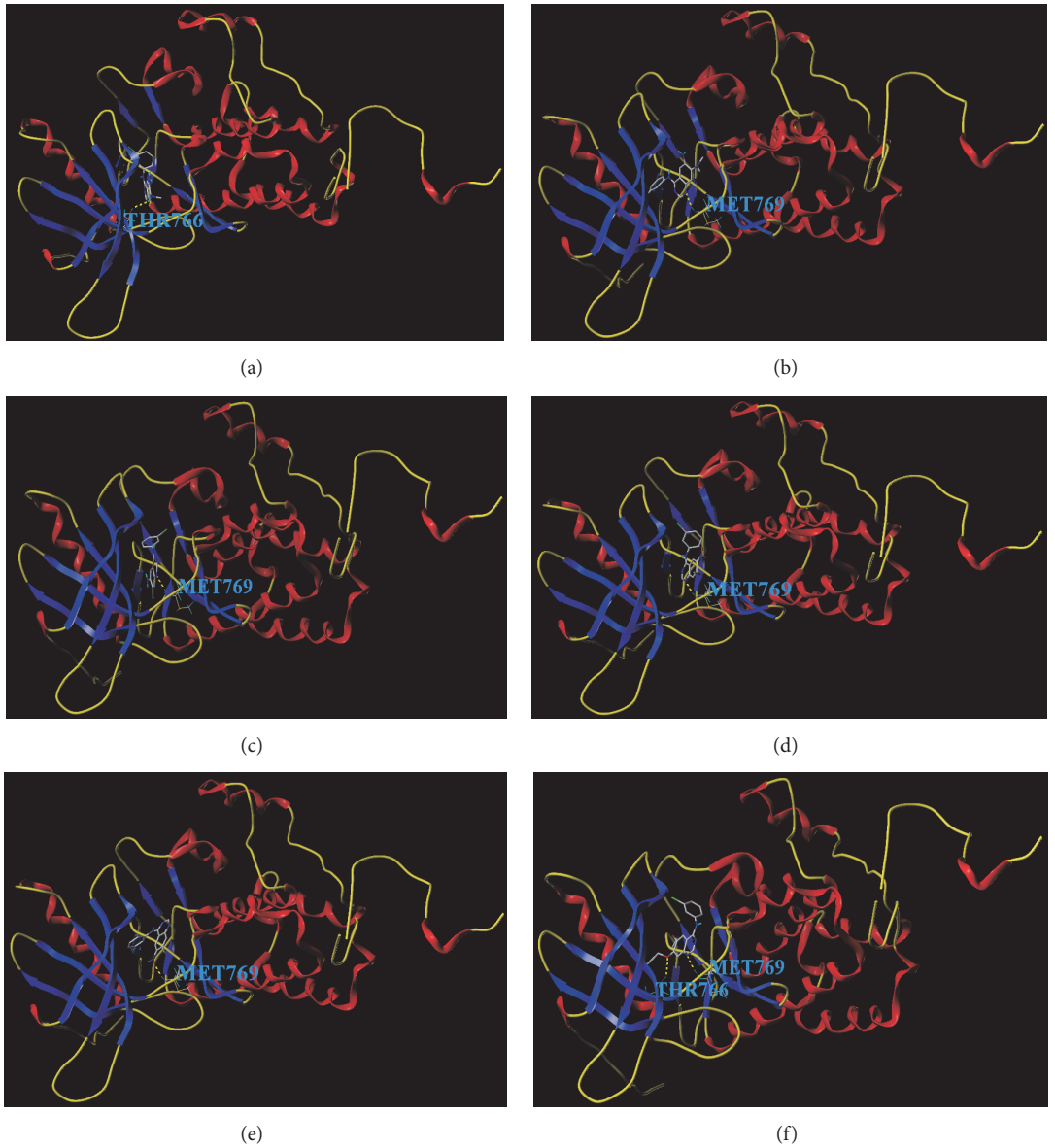
(a) (b) (c) (d)

Figure 3: 3D contour maps of topomer CoMFA model for R1 and R2 of compound **33**. (a) and (c) present steric contour map. (b) and (d) present electrostatic field map. Green, yellow, blue, and red represent large volume, small volume, positively charged, and negatively charged groups, respectively.

(MNE), were employed to analyze the prediction's error. If prediction error is complied with principles I–V defined by Roy, the results were recommended. In our study, the NPE, NNP, AE, AAE, MPE, and MNE were 12, 11, 0.219, 0.526, 0.713, and −0.321, respectively. ABS (MPE/MNE) and R^2 (Y versus residuals) were 2.2 (threshold = 2) and 0.67 (threshold = 0.5), respectively. Hence, it was regarded that our 3D-QSAR model is reliable.

In addition, topomer CoMFA model provides opinions on modifying EGFR inhibitors in order to design potential highly selective and highly active EGFR inhibitors. Compound **33** (see Figure 5) was chosen to study the effect of R1 and R2 groups on activity due to its high activity. In R1 group, large group with a positive-charge in the yloxyethyl increases the compound's bioactivity (see Figure 3). In R2 group, small groups with a positive-charge in the benzene ring may also increase the compound's bioactivity.

4.3. Molecular Docking Analysis. Molecular docking was applied to predict the interaction sites between compounds and EGFR. As the structure of compound **33** is similar to erlotinib, EGFR also interacts with compound **33** at Thr766 and Met769 [50]. Interestingly, it is observed that the binding modes of compound 33-EGFR and erlotinib-EGFR were different despite the similar structure after calculation. Quinolone ring of erlotinib competitively binds to the hydrophobic pocket of EGFR kinase. For erlotinib, the aniline group reached into the pocket, and substituent groups of site 6 and site 7 were located outside of the hydrophobic pocket. For compound **33**, it interacts with the EGFR by substituent groups of site 6 and site 7 in the hydrophobic pocket. In the steric and electrostatic fields, large volume group and positively charged group in site 6 and site 7 of compound **33** may increase inhibitor activity (see Figure 3). Then, the similar chemical series of compound **33** was selected to study



(a) (b) (c) (d) (e) (f)

Figure 4: The docking result of the EGFR inhibitors with EGFR. (a) The binding site of compound **27** with EGFR is Thr766. (b) The binding site of compound **28** with EGFR is Met769. (c) The binding site of compound **30** with EGFR is Met769. (d) The binding site of compound **31** with EGFR is Met769. (e) The binding site of compound **32** with EGFR is Met769. (f) The binding sites of compound **33** with EGFR is Thr766 and Met769.

the docking site. As a result, compounds **28**, **30**, **31**, and **32** interact with EGFR at Met769, and compound **27** interacts with EGFR at Thr766. Thus, we considered that the Thr766 and Met769 played a crucial role in the EGFR activity.

Many studies performed the QSAR on kinase inhibitors, and the result was helpful for drugs design. In Farghaly et al.'s study [60], QSAR model was built, and the RMSE and r^2 were applied to evaluate the model. After calculating, they selected out three predominant descriptors affecting the anticancer activity, and five anticancer agents were screened finally. Sharma showed the 2D-QSAR studies of c-Src tyrosine kinase inhibitors with $q^2 = 0.755$ and $r^2 = 0.832$ [61]. Sharma et al. reported QSAR studies of Aurora A kinase inhibitors [62]. q^2 is 0.762 and r^2 is 0.806. The difference in the number of samples causes the difference in q^2 and r^2. When q^2 and r^2 are more than 0.5 and 0.8, respectively, the model has statistical significance. In our QSAR study, q^2 is 0.565 lower than these two studies, but r^2 is higher (see Table 4). In addition, steric and electrostatic field and molecular docking analysis were applied in our study to explore the activity development and predict the interaction between inhibitors and protein, which is not showed in these studies. In conclusion, QSAR combined with molecular docking provides better insight into the future design of more potent EGFR inhibitors prior to synthesis.

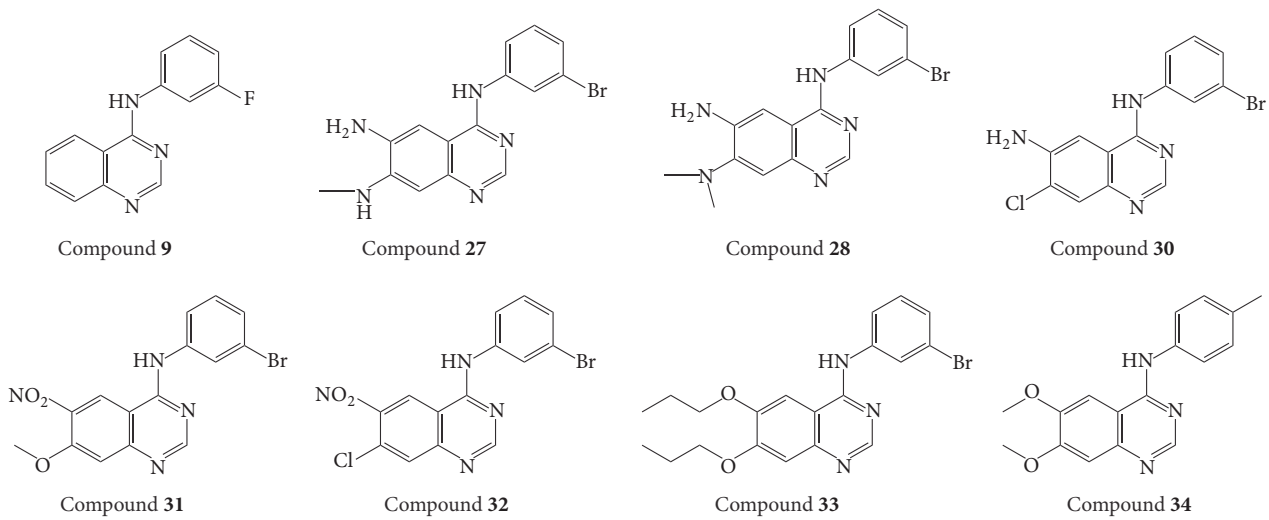


Figure 5: Structures of compounds **9**, **27**, **28**, **30**, **31**, **32**, **33,** and **34**.

Table 4: The comparison of metrics between other studies and ours in QSAR study of the kinase inhibitors.

Metric	QSAR study		
	c-src tyrosine kinase inhibitors [61]	Aurora inhibitors [62]	Our study
q^2	0.755	0.762	0.565
r^2	0.832	0.806	0.888

5. Conclusion

In this study, 2D-QSAR and 3D-QSAR prediction models were built to analyze EGFR inhibitors. Firstly, the 2D-QSAR model was built to predict whether a compound was an inhibitor or a noninhibitor. The accuracy of the 2D-QSAR model using the tenfold cross-validation test and independent set test was 98.99% and 97.67%, respectively. Then, the topomer CoMFA model was built based on EGFR inhibitors. Two models were obtained by cutting different molecular bonds. As a result, model 2 with higher q^2 value and r^2 values was selected to predict EGFR inhibitors. Finally, a series of similar chemical inhibitors were selected to study the interacting sites between EGFR and EGFR inhibitors using molecular docking tool. As a result, Thr766 and Met769 were received by studying the docking result. Thus, we considered that Thr766 and Met769 played a crucial role in the EGFR activity.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Manman Zhao, Lin Wang, and Linfeng Zheng contributed equally to this work.

Acknowledgments

The authors would like to express gratitude towards scholarship from Natural Science Foundation of Shanghai Science and Technology Commission (no. 17ZR1422500), the Shanghai Jiao Tong University Medical Engineering Crossover Fund Project (no. YG2016MS26), the Shanghai University High Performance Computing, Shanghai Municipal Education Commission, the National Natural Science Foundation of China (81271384, 81371623, 31571171, and 31100838), Shanghai Key Laboratory of Bio-Energy Crops (13DZ2272100), the Shanghai Natural Science Foundation (Grant no. 15ZR1414900), the Key Laboratory of Medical Electrophysiology (Southwest Medical University) of Ministry of Education of China (Grant no. 201502), and the Young Teachers of Shanghai Universities Training Program. They also would like to thank Professor Mingyue Zheng from the State Laboratory of Drug Research of Chinese Academy of Science for helping calculate the molecular descriptors.

References

[1] F. Ciardiello and G. Tortora, "EGFR antagonists in cancer treatment," *The New England Journal of Medicine*, vol. 358, no. 11, pp. 1160–1174, 2008.

[2] Y. He, B. S. Harrington, and J. D. Hooper, "New crossroads for potential therapeutic intervention in cancer—intersections between CDCP1, EGFR family members and downstream signaling pathways," *Oncoscience*, vol. 3, no. 1, pp. 5–8, 2016.

[3] R. Wang, X. Wang, J. Q. Wu et al., "Efficient porcine reproductive and respiratory syndrome virus entry in MARC-145 cells requires EGFR-PI3K-AKT-LIMK1-COFILIN signaling pathway," *Virus Research*, vol. 225, pp. 23–32, 2016.

[4] F. Imamura, J. Uchida, Y. Kukita et al., "Monitoring of treatment responses and clonal evolution of tumor cells by circulating tumor DNA of heterogeneous mutant EGFR genes in lung cancer," *Lung Cancer*, vol. 94, pp. 68–73, 2016.

[5] K. Wang, D. Li, and L. Sun, "High levels of EGFR expression in tumor stroma are associated with aggressive clinical features in

epithelial ovarian cancer," *OncoTargets and Therapy*, vol. 9, pp. 377–386, 2016.

[6] A. Cho, J. Hur, Y. W. Moon et al., "Correlation between EGFR gene mutation, cytologic tumor markers, 18F-FDG uptake in non-small cell lung cancer," *BMC Cancer*, vol. 16, no. 1, article 224, 2016.

[7] X. B. Holdman, T. Welte, K. Rajapakshe et al., "Upregulation of EGFR signaling is correlated with tumor stroma remodeling and tumor recurrence in FGFR1-driven breast cancer," *Breast Cancer Research*, vol. 17, no. 1, article 141, 2015.

[8] C. Sarkar, "Epidermal growth factor receptor (EGFR) gene amplification in high grade gliomas," *Neurology India*, vol. 64, no. 1, pp. 27-28, 2016.

[9] K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano, "A comprehensive pathway map of epidermal growth factor receptor signaling," *Molecular Systems Biology*, vol. 1, p. E1, 2005.

[10] J. C. Baer, A. A. Freeman, E. S. Newlands, A. J. Watson, J. A. Rafferty, and G. P. Margison, "Depletion of O^6-alkylguanine-DNA alkyltransferase correlates with potentiation of temozolomide and CCNU toxicity in human tumour cells," *British Journal of Cancer*, vol. 67, no. 6, pp. 1299–1302, 1993.

[11] N. Minkovsky and A. Berezov, "BIBW-2992, a dual receptor tyrosine kinase inhibitor for the treatment of solid tumors," *Current Opinion in Investigational Drugs*, vol. 9, no. 12, pp. 1336–1346, 2008.

[12] J. C. Dearden, "The history and development of quantitative structure-activity relationships (QSARs)," *International Journal of Quantitative Structure-Property Relationships*, vol. 1, no. 1, pp. 1–44, 2016.

[13] A. Cherkasov, E. N. Muratov, D. Fourches et al., "QSAR modeling: where have you been? Where are you going to?" *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[14] K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, 2015.

[15] D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti, and O. Nicolotti, "Applicability domain for QSAR models: where theory meets reality," *International Journal of Quantitative Structure-Property Relationships*, vol. 1, no. 1, pp. 45–63, 2016.

[16] K. Roy, S. Kar, and P. Ambure, "On a simple approach for determining applicability domain of QSAR models," *Chemometrics and Intelligent Laboratory Systems*, vol. 145, pp. 22–29, 2015.

[17] S. Lee and M. G. Barron, "Development of 3D-QSAR model for acetylcholinesterase inhibitors using a combination of fingerprint, molecular docking, and structure-based pharmacophore approaches," *Toxicological Sciences*, vol. 148, no. 1, pp. 60–70, 2015.

[18] G. Tresadern and D. Bemporad, "Modeling approaches for ligand-based 3D similarity," *Future Medicinal Chemistry*, vol. 2, no. 10, pp. 1547–1561, 2010.

[19] K.-C. Chou, "Structural bioinformatics and its impact to biomedical science," *Current Medicinal Chemistry*, vol. 11, no. 16, pp. 2105–2134, 2004.

[20] J.-W. Liang, T.-J. Zhang, Z.-J. Li, Z.-X. Chen, X.-L. Yan, and F.-H. Meng, "Predicting potential antitumor targets of Aconitum alkaloids by molecular docking and protein–ligand interaction fingerprint," *Medicinal Chemistry Research*, vol. 25, no. 6, pp. 1115–1124, 2016.

[21] L. Blake and M. E. S. Soliman, "Identification of irreversible protein splicing inhibitors as potential anti-TB drugs: insight from hybrid non-covalent/covalent docking virtual screening and molecular dynamics simulations," *Medicinal Chemistry Research*, vol. 23, no. 5, pp. 2312–2323, 2014.

[22] P. Ambure, S. Kar, and K. Roy, "Pharmacophore mapping-based virtual screening followed by molecular docking studies in search of potential acetylcholinesterase inhibitors as anti-Alzheimer's agents," *BioSystems*, vol. 116, no. 1, pp. 10–20, 2014.

[23] J. Lv, J. Su, F. Wang, Y. Qi, H. Liu, and Y. Zhang, "Detecting novel hypermethylated genes in Breast cancer benefiting from feature selection," *Computers in Biology and Medicine*, vol. 40, no. 2, pp. 159–167, 2010.

[24] F. R. Ajdadi, Y. A. Gilandeh, K. Mollazade, and R. P. Hasanzadeh, "Application of machine vision for classification of soil aggregate size," *Soil and Tillage Research*, vol. 162, pp. 8–17, 2016.

[25] R. Sadeghia, R. Zarkami, K. Sabetraftar, and P. Van Damme, "Application of genetic algorithm and greedy stepwise to select input variables in classification tree models for the prediction of habitat requirements of Azolla filiculoides (Lam.) in Anzali wetland, Iran," *Ecological Modelling*, vol. 251, pp. 44–53, 2013.

[26] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[27] F. Imani, F. E. Boada, F. S. Lieberman, D. K. Davis, and J. M. Mountz, "Molecular and metabolic pattern classification for detection of brain glioma progression," *European Journal of Radiology*, vol. 83, no. 2, pp. e100–e105, 2014.

[28] K. E. Emblem, F. G. Zoellner, B. Tennoe et al., "Predictive modeling in glioma grading from MR perfusion images using support vector machines," *Magnetic Resonance in Medicine*, vol. 60, no. 4, pp. 945–952, 2008.

[29] R. D. Cramer, "Topomer CoMFA: a design methodology for rapid lead optimization," *Journal of Medicinal Chemistry*, vol. 46, no. 3, pp. 374–388, 2003.

[30] K. Z. Myint and X.-Q. Xie, "Recent advances in fragment-based QSAR and multi-dimensional QSAR methods," *International Journal of Molecular Sciences*, vol. 11, no. 10, pp. 3846–3866, 2010.

[31] C. G. Gadhe, "CoMFA vs. Topomer CoMFA, which one is better a case study with 5-lipoxygenase inhibitors," *Journal of the Chosun Natural Science*, vol. 4, no. 2, pp. 91–98, 2011.

[32] G. W. Rewcastle, W. A. Denny, A. J. Bridges et al., "Tyrosine kinase inhibitors. 5. Synthesis and structure-activity relationships for 4-[(phenylmethyl)amino]- and 4-(phenylamino)quinazolines as potent adenosine 5′-triphosphate binding site inhibitors of the tyrosine kinase domain of the epidermal growth factor receptor," *Journal of Medicinal Chemistry*, vol. 38, no. 18, pp. 3482–3487, 1995.

[33] A. M. Thompson, A. J. Bridges, D. W. Fry, A. J. Kraker, and W. A. Denny, "Tyrosine kinase inhibitors. 7. 7-Amino-4-(phenylamino)- and 7-amino-4-[(phenylmethyl)amino]pyrido[4,3-d]pyrimidines: a new class of inhibitors of the tyrosine kinase activity of the epidermal growth factor receptor," *Journal of Medicinal Chemistry*, vol. 38, no. 19, pp. 3780–3788, 1995.

[34] G. W. Rewcastle, A. J. Bridges, D. W. Fry, J. R. Rubin, and W. A. Denny, "Tyrosine kinase inhibitors. 12. Synthesis and structure-activity relationships for 6-substituted 4-(phenylamino)pyrimido[5,4-d]pyrimidines designed as inhibitors of the epidermal growth factor receptor," *Journal of Medicinal Chemistry*, vol. 40, no. 12, pp. 1820–1826, 1997.

[35] A. M. Thompson, D. K. Murray, W. L. Elliott et al., "Tyrosine kinase inhibitors. 13. Structure—activity relationships for

soluble 7-substituted 4-[(3-bromophenyl)amino]pyrido[4,3-d]pyrimidines designed as inhibitors of the tyrosine kinase activity of the epidermal growth factor receptor," *Journal of Medicinal Chemistry*, vol. 40, no. 24, pp. 3915–3925, 1997.

[36] A. J. Bridges, H. Zhou, D. R. Cody et al., "Tyrosine kinase inhibitors. 8. An unusually steep structure-activity relationship for analogues of 4-(3-bromoanilino)-6,7-dimethoxyquinazoline (PD 153035), a potent inhibitor of the epidermal growth factor receptor," *Journal of Medicinal Chemistry*, vol. 39, no. 1, pp. 267–276, 1996.

[37] G. W. Rewcastle, B. D. Palmer, A. M. Thompson et al., "Tyrosine kinase inhibitors. 10. Isomeric 4-[(3-bromophenyl)amino]pyrido[d]-pyrimidines are potent ATP binding site inhibitors of the tyrosine kinase function of the epidermal growth factor receptor," *Journal of Medicinal Chemistry*, vol. 39, no. 9, pp. 1823–1835, 1996.

[38] S. Li, C. Guo, H. Zhao, Y. Tang, and M. Lan, "Synthesis and biological evaluation of 4-[3-chloro-4-(3-fluorobenzyloxy) anilino]-6-(3-substituted-phenoxy)pyrimidines as dual EGFR/ErbB-2 kinase inhibitors," *Bioorganic and Medicinal Chemistry*, vol. 20, no. 2, pp. 877–885, 2012.

[39] A. G. Waterson, K. G. Petrov, K. R. Hornberger et al., "Synthesis and evaluation of aniline headgroups for alkynyl thienopyrimidine dual EGFR/ErbB-2 kinase inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 19, no. 5, pp. 1332–1336, 2009.

[40] N. Suzuki, T. Shiota, F. Watanabe et al., "Synthesis and evaluation of novel pyrimidine-based dual EGFR/Her-2 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 21, no. 6, pp. 1601–1606, 2011.

[41] N. Suzuki, T. Shiota, F. Watanabe et al., "Discovery of novel 5-alkynyl-4-anilinopyrimidines as potent, orally active dual inhibitors of EGFR and Her-2 tyrosine kinases," *Bioorganic and Medicinal Chemistry Letters*, vol. 22, no. 1, pp. 456–460, 2012.

[42] J. J. Irwin, "Software review: ChemOffice 2005 Pro by Cambridgesoft," *Journal of Chemical Information and Modeling*, vol. 45, no. 5, pp. 1468-1469, 2005.

[43] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, supplement 5, pp. 91–98, 2006.

[44] L. Wang, H. Shen, B. Li, and D. Hu, "Classification of schizophrenic patients and healthy controls using multiple spatially independent components of structural MRI data," *Frontiers of Electrical and Electronic Engineering in China*, vol. 6, no. 2, pp. 353–362, 2011.

[45] Y. P. Zhang, N. Sussman, G. Klopman, and H. S. Rosenkranz, "Development of methods to ascertain the predictivity and consistency of SAR models: application to the U.S. National toxicology program rodent carcinogenicity bioassays," *Quantitative Structure-Activity Relationships*, vol. 16, no. 4, pp. 290–295, 1997.

[46] K. Roy, R. N. Das, P. Ambure, and R. B. Aher, "Be aware of error measures. Further studies on validation of predictive QSAR models," *Chemometrics and Intelligent Laboratory Systems*, vol. 152, pp. 18–33, 2016.

[47] S. Yu, J. Yuan, J. Shi et al., "HQSAR and topomer CoMFA for predicting melanocortin-4 receptor binding affinities of trans-4-(4-chlorophenyl) pyrrolidine-3-carboxamides," *Chemometrics and Intelligent Laboratory Systems*, vol. 146, pp. 34–41, 2015.

[48] S. Kumar and M. Tiwari, "Topomer-CoMFA-based predictive modelling on 2,3-diaryl-substituted-1,3-thiazolidin-4-ones as non-nucleoside reverse transcriptase inhibitors," *Medicinal Chemistry Research*, vol. 24, no. 1, pp. 245–257, 2015.

[49] Y. Tian, Y. Shen, X. Zhang et al., "Design some new type-I c-met inhibitors based on molecular docking and topomer comfa research," *Molecular Informatics*, vol. 33, no. 8, pp. 536–543, 2014.

[50] G. Tresadern, J.-M. Cid, and A. A. Trabanco, "QSAR design of triazolopyridine mGlu2 receptor positive allosteric modulators," *Journal of Molecular Graphics and Modelling*, vol. 53, pp. 82–91, 2014.

[51] H. Tang, L. Yang, J. Li, and J. Chen, "Molecular modelling studies of 3,5-dipyridyl-1,2,4-triazole derivatives as xanthine oxidoreductase inhibitors using 3D-QSAR, Topomer CoMFA, molecular docking and molecular dynamic simulations," *Journal of the Taiwan Institute of Chemical Engineers*, vol. 68, pp. 64–73, 2016.

[52] S. D. Joshi, U. A. More, D. Koli, M. S. Kulkarni, M. N. Nadagouda, and T. M. Aminabhavi, "Synthesis, evaluation and in silico molecular modeling of pyrroyl-1,3,4-thiadiazole inhibitors of InhA," *Bioorganic Chemistry*, vol. 59, pp. 151–167, 2015.

[53] J. Stamos, M. X. Sliwkowski, and C. Eigenbrot, "Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor," *Journal of Biological Chemistry*, vol. 277, no. 48, pp. 46265–46272, 2002.

[54] J. Miao and L. Niu, "A survey on feature selection," *Procedia Computer Science*, vol. 91, pp. 919–926, 2016.

[55] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," *Progress in Artificial Intelligence*, vol. 5, no. 2, pp. 65–75, 2016.

[56] B. Niu, Q. Su, X. Yuan, W. Lu, and J. Ding, "QSAR study on 5-lipoxygenase inhibitors based on support vector machine," *Medicinal Chemistry*, vol. 8, no. 6, pp. 1108–1116, 2012.

[57] W. Ding, M. Sun, S. Luo et al., "A 3D QSAR study of betulinic acid derivatives as anti-tumor agents using topomer CoMFA: model building studies and experimental verification," *Molecules*, vol. 18, no. 9, pp. 10228–10241, 2013.

[58] Y. Xiang, J. Song, and Z. Zhang, "Topomer CoMFA and virtual screening studies of azaindole class renin inhibitors," *Combinatorial Chemistry and High Throughput Screening*, vol. 17, no. 5, pp. 458–472, 2014.

[59] K. Roy, P. Ambure, and R. B. Aher, "How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models?" *Chemometrics and Intelligent Laboratory Systems*, vol. 162, pp. 44–54, 2017.

[60] T. A. Farghaly, H. M. E. Hassaneen, and H. S. A. Elzahabi, "Eco-friendly synthesis and 2D-QSAR study of novel pyrazolines as potential anticolon cancer agents," *Medicinal Chemistry Research*, vol. 24, no. 2, pp. 652–668, 2015.

[61] M. C. Sharma, "2D QSAR studies of the inhibitory activity of a series of substituted purine derivatives against c-Src tyrosine kinase," *Journal of Taibah University for Science*, vol. 10, no. 4, pp. 563–570, 2016.

[62] M. C. Sharma, S. Sharma, and K. Bhadoriya, "QSAR studies on pyrazole-4-carboxamide derivatives as Aurora A kinase inhibitors," *Journal of Taibah University for Science*, vol. 10, no. 1, pp. 107–114, 2016.

Hindawi
BioMed Research International
Volume 2017, Article ID 1645619, 6 pages
https://doi.org/10.1155/2017/1645619

Research Article

A Cancer Gene Selection Algorithm Based on the K-S Test and CFS

Qiang Su,[1] Yina Wang,[2] Xiaobing Jiang,[3] Fuxue Chen,[4] and Wen-cong Lu[5]

[1]*School of Communication & Information Engineering, Shanghai University, Shanghai 2000444, China*
[2]*Department of VIP Medical Center, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China*
[3]*Department of Neurosurgery, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, No. 651, Dongfeng Road E, Guangzhou 510060, China*
[4]*College of Life Sciences, Shanghai University, Shanghai 2000444, China*
[5]*Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Fuxue Chen; chenfuxue@staff.shu.edu.cn and Wen-cong Lu; wclu@shu.edu.cn

Received 12 January 2017; Accepted 6 April 2017; Published 8 May 2017

Academic Editor: Jianxin Li

Background. To address the challenging problem of selecting distinguished genes from cancer gene expression datasets, this paper presents a gene subset selection algorithm based on the Kolmogorov-Smirnov (K-S) test and correlation-based feature selection (CFS) principles. The algorithm selects distinguished genes first using the K-S test, and then, it uses CFS to select genes from those selected by the K-S test. *Results*. We adopted support vector machines (SVM) as the classification tool and used the criteria of accuracy to evaluate the performance of the classifiers on the selected gene subsets. This approach compared the proposed gene subset selection algorithm with the K-S test, CFS, minimum-redundancy maximum-relevancy (mRMR), and ReliefF algorithms. The average experimental results of the aforementioned gene selection algorithms for 5 gene expression datasets demonstrate that, based on accuracy, the performance of the new K-S and CFS-based algorithm is better than those of the K-S test, CFS, mRMR, and ReliefF algorithms. *Conclusions*. The experimental results show that the K-S test-CFS gene selection algorithm is a very effective and promising approach compared to the K-S test, CFS, mRMR, and ReliefF algorithms.

1. Introduction

Big data analysis technology can mine gene information related to diseases and drugs from massive gene data and provide new ideas for drug development as well as disease diagnosis and treatment. Therefore, big data has positive effects on cancer research. Genetic data analysis includes four steps: gene data acquisition, gene data pretreatment, gene selection, and classification model establishment and evaluation. Of these steps, genetic data acquisition is a biomedical process, and the other steps are data mining processes. This paper focuses on the gene selection step in genetic data analysis by exploring the challenges to gene data analysis and effective strategies and methods for gene selection.

According to its relationship with the classifier, the feature (gene) selection method is divided into the filter method, the wrapper method, and the embedded method. The filter method selects the features that contribute to the classification, which is independent of the learning process, and has a higher efficiency and a stronger generalization ability. The wrapper method selects the corresponding feature subsets according to the classification performance of the feature subsets. Depending on the learning process, the wrapper method has a higher accuracy, but it is prone to overadaptability, a poor generalization performance, and low time efficiency. The combination of the filter method and the wrapper method is a new trend in studies of feature selection.

There is a significant difference in the expression value of a discriminative gene between different genotypes. Thus, a series of filter-based gene selection methods, based on parametric statistics, was developed to detect whether there were significant differences between genotypes and to select a subset of genes with significant differences [1, 2]. However,

Table 1: Dataset.

Dataset	Samples	Genes
Breast cancer	97	24481
Lung cancer	181	12533
Colon tumor	62	2000
Ovarian cancer	253	15154
Leukemia	72	7129

parametric statistical methods need to assume a Gaussian distribution of the data, and the actual genetic dataset usually does not meet the Gaussian distribution hypothesis. Therefore, a nonparametric statistical method, the Wilcoxon rank sum test, is used in gene selection studies. However, the rank sum test can be used to reveal the location of two sample types (the distributions of the values of the two sample types) only when the sample size is large or the measurement level is low (the sample observations have only a small number of values). When the sample size is very small or has the same rank value as the sample with the same rank, it is not appropriate to use the rank sum test for gene selection.

The Kolmogorov-Smirnov (K-S) test is another nonparametric statistical method used to compare the distribution of two sample types. This method is very sensitive to the difference of the distribution of two sample types. It has been successfully applied in the analysis of ovarian cancer gene data, recognition, and other fields [3]. However, an independent nonparametric test method does not take into account the redundancy of the genes in the selection of genes with discriminatory power.

The correlation-based feature selection (CFS) [4, 5] method can efficiently select subsets of genes that are highly correlated with the class and that have low redundancy. However, due to the high-dimensional characteristics of gene datasets, it is very time-consuming to adopt the CFS method for gene selection directly. Therefore, a gene selection algorithm combining the K-S test and CFS is proposed in this paper. Most of the redundant and noise genes are removed by the K-S test, and the genes with a significant distinguishing ability are retained. Then, CFS is used to evaluate the genes that are highly correlated with the class and have low redundancy. A support vector machine (SVM) [6, 7] is used as the classifier to evaluate the gene subsets generated based on accuracy. Finally, the method is compared with the K-S test, minimum-redundancy maximum-relevancy (mRMR) [8], and classic ReliefF algorithms [9]. The experimental results from five gene datasets show that the K-S test-CFS gene selection method is an effective gene selection algorithm.

2. Materials and Methods

2.1. Datasets Description. In this paper, five classical cancer gene datasets are used: breast cancer [10], lung cancer [11], colon tumor [12], ovarian cancer [13], and leukemia [2]. Detailed information on the datasets is listed in Table 1. To eliminate the influence of different dimensions on the experimental results, the five datasets were Z-score standardized as part of the preprocessing.

2.2. K-S Test. In this paper, the K-S test was used to determine significant differences between the genes of the tumor patients and those of normal controls. Let $X_1, X_2, X_3, \ldots, X_N$ be a gene X from the gene dataset, and the observed value is $x_1, x_2, \ldots, x_n$, where n is the sample number of the gene dataset. According to the gene order value, the order of the observations is $x_{(1)} \ll x_{(2)} \ll \cdots \ll x_{(n)}$, and the cumulative distribution function of the gene X is defined as follows:

$$F(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} \le x \le x_{(k+1)},\ k = 1, \ldots, n-1 \\ 1, & x > X_{(n)}. \end{cases} \quad (1)$$

Assuming that the cumulative distribution functions of the gene to be tested in the tumor sample and the normal sample are $F_1(x)$ and $F_2(x)$, where the number of observations is the number of positive and negative samples, the K-S test statistic is

$$D = \max_x \left| F_1(x) - F_2(x) \right|. \quad (2)$$

According to the K-S test theory, when $D < D_{\text{crit}}$ (the critical value of D_{crit} for the level of significance α), the gene has no significant difference between the positive and negative classes when the significance level is α; if $D \ge D_{\text{crit}}$, there is a significant difference between the positive and negative samples at the $1 - \alpha$ confidence level.

From (2), we can see that the bigger the D value, the greater the difference between the positive and negative classes of the gene, indicating a stronger ability to distinguish between the positive and negative samples.

2.3. Correlation-Based Feature Selection (CFS). The correlation feature selection (CFS) method evaluates subsets of features according to the following hypothesis: "good feature subsets contain features that are highly correlated with the classification yet uncorrelated to each other." The bias of the evaluation function is towards subsets containing features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they have a low correlation with the class. Redundant features should be removed, as they will be highly correlated with one or more of the remaining features. The acceptance of a feature depends on the extent to which it predicts classes in areas of the instance space not already predicted by other features.

2.4. K-S Test-CFS Method for Gene Selection. As we previously mentioned, the K-S test is a general and successful

Table 2: The number of genes selected by K-S test, Wilcoxon test, and T test in five datasets with different alpha.

Dataset	Algorithm	Alpha = 1	Alpha = 0.05	Alpha = 0.01	Alpha = 0.005	Alpha = 0.001
Breast cancer	K-S	24481	3502	1397	940	349
	Wilcoxon	24481	3829	1529	1029	381
	T	24481	3251	1161	726	273
Lung cancer	K-S	12533	2886	1982	1588	1300
	Wilcoxon	12533	3225	2658	1986	1528
	T	12533	3190	2580	1996	1625
Colon tumor	K-S	2000	324	146	105	44
	Wilcoxon	2000	387	188	140	59
	T	2000	389	171	113	53
Ovarian cancer	K-S	15154	7268	3408	1386	268
	Wilcoxon	15154	7652	3927	1876	329
	T	15154	7900	3848	1938	318
Leukemia	K-S	7129	1716	1036	843	524
	Wilcoxon	7129	1860	1169	962	644
	T	7129	1811	1115	931	583

attribute estimator and is able to effectively provide quality estimates of attributes in problems that have dependencies between attributes. However, the K-S test does not explicitly reduce the redundancy in selected genes. CFS selects genes that have the highest relevance with the target class and that are also maximally dissimilar to each other. Thus, the integration of the K-S test and CFS leads to an effective gene selection scheme.

The details of the K-S test-CFS algorithm are as follows: in the first stage, the K-S test is applied to find a candidate gene set. This approach removes many unimportant genes and reduces the computational load for CFS. In the second stage, the CFS method is applied to directly and explicitly reduce the redundancy and to select a compact yet effective gene subset from the candidate set.

2.5. Software Package. In this paper, the K-S test, T test, and Wilcoxon test algorithms are implemented using MATLAB R2012a. The CFS, mRMR, ReliefF, and SVM algorithms are implemented using Weka 3.6. Weka (http://www.cs.waikato.ac.nz/ml/weka/) is a software packaged that collects various types of learning algorithms for data mining tasks. The SVM algorithm uses a linear kernel function, and the penalty factor C takes a fixed value of 1.

3. Results and Discussion

3.1. Comparison of the K-S Test with the T Test and the Wilcoxon Test. This section compares the performance of the gene selection algorithms using the K-S test, the Wilcoxon test, and the T-test. First, the significance level alpha was set, and then, each gene in the dataset was tested by the K-S test, the Wilcoxon test, and the T test to select the important genes in order to form a subset of preselected genes. In the preselected gene subset, SVM was used as the classifier to calculate the accuracy of the 10-fold cross-validation. Then, a performance comparison of the gene subsets selected by the K-S test, the Wilcoxon test, and the T test in the different alpha values was performed. Table 2 lists the number of gene subsets selected by the K-S test, the Wilcoxon test, and the T test in the five datasets with different alpha values. Table 3 shows the average classification accuracy of the 10-fold cross-validation in the gene subsets selected by the K-S test, the Wilcoxon test, and the T-test in the five datasets with different alpha values.

The experimental results in Table 2 show that the number of gene subsets selected by the K-S test, the Wilcoxon test, and the T test with the same alpha value was different. As shown in Table 2, the K-S test selected a smaller subset of genes in most cases.

Table 2 also shows that the subset of genes selected by the three test algorithms was smaller when the confidence level was large and the significance level α was small. When the confidence level was 99.9%, the significance level $\alpha = 0.001$. In the colon dataset, the size of the selected subset of genes was approximately 50, which is approximately 2.5% of the original dataset. The size of the subset of genes selected in the breast cancer dataset was approximately 1.5% of the original number of genes in the dataset. The worst case observed was with the lung cancer dataset, and at this significant level, the size of the selected gene subset for the three test algorithms was approximately 10% of the original gene number of genes in the dataset.

The above analysis shows that the K-S test is a very effective genetic importance measurement algorithm. This test selected a smaller subset of genes that had a high interclass discrimination ability.

The average classification accuracy of the subset of genes selected by the three test algorithms at the different levels of significance is shown in Table 3. For the breast cancer dataset, the significance level was 0.001, and the average classification accuracy rate of the K-S test was slightly worse than that of the Wilcoxon test; however it was better than that of the T test. When the significance level was 0.0.5, 0.01, or 0.005,

Table 3: The average classification accuracy (%) of 10-fold cross-validation in the gene subsets selected by K-S, Wilcoxon test, and T-test in five datasets with different alpha.

Dataset	Algorithm	Alpha = 1	Alpha = 0.05	Alpha = 0.01	Alpha = 0.005	Alpha = 0.001
Breast cancer	K-S	68.6	83.6	86.3	86.3	83.2
	Wilcoxon	67.8	83.5	84.8	84.8	84.8
	T	66.7	80.2	83.5	80.5	80.5
Lung cancer	K-S	85.8	89.6	90.4	91.6	91.6
	Wilcoxon	85.8	86.9	88.5	89.5	89.5
	T	83.6	86.9	88.5	89.5	89.5
Colon tumor	K-S	73.4	81.4	85.9	85.9	83.2
	Wilcoxon	73.4	79.2	80.2	81.5	83.2
	T	73.4	79.2	80.2	81.5	81.5
Ovarian cancer	K-S	95.3	98.6	100	100	96.5
	Wilcoxon	95.3	97.3	98.6	100	94.6
	T	95.3	97.3	98.6	100	94.6
Leukemia	K-S	71.6	75.3	81.4	82.6	85.6
	Wilcoxon	71.6	75.3	81.4	81.4	83.5
	T	71.6	75.3	81.4	81.4	82.2

Table 4: The comparisons in CFS, mRMR, and ReliefF algorithms.

Dataset	Gene selection method					
	CFS		mRMR		ReliefF	
	The number of genes	Accuracy	The number of genes	Accuracy	The number of genes	Accuracy
Breast cancer	11.7	87.4	10.4	85.6	15.9	59.5
Lung cancer	23.2	91.6	25.7	88.4	26.7	87.6
Colon tumor	10.7	90.1	12.6	86.4	15.3	84.8
Ovarian cancer	33.2	98.5	31.5	95.6	37.4	93.2
Leukemia	25.2	99.6	2.5	99.6	16.4	77.6

the average classification accuracy rate of the K-S test was not lower than the rates of the Wilcoxon test and the T test. For the other four genetic datasets, regardless of whether the significance level was 0.05, 0.01, 0.005, or 0.001, the average classification accuracy rate of the gene subset selected by the K-S test was not lower than the rates of the Wilcoxon test and the T test. Therefore, this finding demonstrated that the K-S test could select a better gene subset.

Based on the above results, the K-S test was superior to the Wilcoxon test and the T test for gene selection.

3.2. Compare the CFS with the mRMR and ReliefF Algorithms. The CFS algorithm was compared to the mRMR and ReliefF algorithms to validate the performance of the gene selection in the preselected gene subset. First, all of the genes were prescreened by the K-S test with a significance level of 0.01, and a preselected gene subset was obtained. The CFS algorithm selected the appropriate subset of genes directly from the subset of prescreened genes. The mRMR and ReliefF algorithms selected the first 50 genes sorted by the importance of the gene. Then, a forward selection algorithm was used to select the appropriate subset of genes from those 50 genes.

In the experiment, we adopted SVM as a classifier and used the criteria of the average accuracy of a tenfold cross-validation in the dataset to evaluate the performance of the classifiers on the selected gene subsets. To obtain statistically significant experimental results, the dataset samples were randomly shuffled, the procedure was repeated 10 times, and the average of the 10 replicates was recorded and compared. Table 4 shows the average accuracy of the tenfold cross-validation of the three algorithms in the five gene datasets and the corresponding number of genes on average.

From the comparison of the average accuracy (calculated from the results of the ten replicates) of the three algorithms shown in Table 4, we can see that, for the breast cancer dataset, the CFS algorithm achieves the best performance with the least features, which is significantly better than the performance of the other algorithms. For the colon dataset, the CFS was superior to the ReliefF and mRMR algorithms. For the lung cancer, ovarian, and leukemia datasets, the performance of the CFS algorithm is similar to that of the mRMR algorithm and better than that of the ReliefF algorithm.

Based on the above results, the CFS algorithm is superior to the mRMR and ReliefF algorithms for the preselected gene subset.

Table 5: The comparisons in K-S test-CF, K-S, CFS, mRMR, and ReliefF algorithms.

Dataset	Gene selection method									
	k-S test- CFS		CFS		k-S test		mRMR		ReliefF	
	The number of genes	Accuracy	The number of genes	Accuracy	The number of genes	Accuracy	The number of genes	Accuracy	The number of genes	Accuracy
Breast cancer	11.7	87.4	19.6	80.5	22.5	78.8	21.8	82.4	15.9	59.4
Lung cancer	23	91.6	27.3	88.9	33.4	80.6	289	89.8	33.6	84.7
Colon tumor	10.7	90.1	6.8	89.7	19.4	84.5	5.9	89.7	15	74.9
Ovarian cancer	33.2	98.5	31.6	95.3	46	78.9	32.7	95.2	39.6	90.6
Leukemia	25.2	79.6	33.3	78.9	38.7	72.7	28.6	75.7	36.4	77.6

3.3. Comparison of the K-S Test-CF Algorithm with the K-S Test, CFS, mRMR, and ReliefF Algorithms. We also compared the K-S test-CFS selection algorithm with other gene selection algorithms, including the K-S test, mRMR, CFS, and ReliefF. Table 5 presents the classification accuracy comparison using the SVM classifier based on the selected genes and these five feature selection methods. From Table 5, we observed the following:

(i) The K-S test-CFS algorithm achieved a better performance than the other gene selection algorithms on almost all datasets. The experimental comparisons demonstrate the effectiveness of the integration of the K-S test and CFS.

(ii) CFS achieved a good performance on most of the datasets. However, its performance was not always as good as that of the K-S test-CFS algorithm. It outperforms the mRMR and ReliefF algorithms.

In summary, the performance of the K-S test-CFS is superior to other gene filtering algorithms. However, in the course of the experiment, we found that the runtime of the K-S test-CFS had no advantage over the other algorithms. Therefore, the focus of the next step in this work should be how to optimize the running time of the K-S test-CFS algorithm.

4. Conclusions

In this paper, we present a K-S test-CFS selection algorithm developed by combining the K-S test and CFS. The K-S test effectively provided quality estimates of the attributes in problems that have dependencies between attributes, and the CFS method selected genes that had the highest relevance with the target class and are also maximally dissimilar to each other. The integration of the K-S test and CFS thus leads to an effective gene selection scheme. In the first stage, the K-S test is applied to find a candidate gene set. In the second stage, CFS is applied to select a compact yet effective gene subset from the candidate set. Comprehensive experiments were conducted to compare the K-S test-CFS selection algorithm to the K-S test, CFS, ReliefF, and mRMR feature selection methods using the SVM classifier on five different datasets. The experimental results show that the K-S test-CFS gene selection is an effective method compared to the K-S test, CFS, mRMR, and ReliefF algorithms.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Qiang Su and Yina Wang contributed equally to this work. Fuxue Chen and Wencong Lu conceived the project. Qiang Su designed the methodology, performed the experiments, and interpreted the results, and Xiaobing Jiang drafted the manuscript. Yina Wang revised the manuscript.

Acknowledgments

The present study was supported by The National Key Research and Development Program of China (Grant no. 2016YFD0501101), National Natural Science Foundation of China (81271384 and 81371623), and High Performance Computing Center Program of Shanghai University.

References

[1] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[3] J. S. Yu, S. Ongarello, R. Fiedler et al., "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data," *Bioinformatics*, vol. 21, no. 10, pp. 2200–2209, 2005.

[4] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning*, The University of Waikato, Hamilton, New Zealand, 1999.

[5] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast Correlation Based Filter (FCBF) with a different search strategy," in *Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS '08)*, Istanbul Technical University, Suleyman Demirel Cultural Center, Istanbul, Turkey, October 2008.

[6] A. Ben-Hur, D. Horn, and H. T. Siegelmann, "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125–137, 2001.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[9] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[10] L. J. Van't Veer, H. Dai, M. J. Van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[11] G. J. Gordon, R. V. Jensen, L.-L. Hsiao et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[12] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[13] E. F. Petricoin, A. M. Ardekani, B. A. Hitt et al., "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, no. 9306, pp. 572–577, 2002.

Hindawi
BioMed Research International
Volume 2017, Article ID 4751260, 10 pages
https://doi.org/10.1155/2017/4751260

Research Article

Curcumin Analogue CA15 Exhibits Anticancer Effects on HEp-2 Cells via Targeting NF-κB

Jian Chen,[1,2] Linlin Zhang,[3] Yilai Shu,[1] Liping Chen,[2] Min Zhu,[2] Song Yao,[2] Jiabing Wang,[2] Jianzhang Wu,[2] Guang Liang,[2] Haitao Wu,[1] and Wulan Li[2,4]

[1]*Departments of Otolaryngology-Head and Neck Surgery, Eye, Ear, Nose and Throat Hospital, Fudan University, Shanghai, China*
[2]*Chemical Biology Research Center, College of Pharmaceutical Sciences, Wenzhou Medical University, Wenzhou, Zhejiang, China*
[3]*Departments of Stomatology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, Zhejiang, China*
[4]*College of Information Science and Computer Engineering, School of the first Clinical Medical Sciences, Wenzhou Medical University, Wenzhou, Zhejiang, China*

Correspondence should be addressed to Haitao Wu; eentwuhaitao@163.com and Wulan Li; lwlwzmu@163.com

Received 7 October 2016; Revised 20 February 2017; Accepted 26 February 2017; Published 20 March 2017

Academic Editor: Bing Niu

Laryngeal carcinoma remains one of the most common malignancies, and curcumin has been proven to be effective against head and neck cancers in vitro. However, it has not yet been applied in clinical settings due to its low stability. In the current study, we synthesized 34 monocarbonyl analogues of curcumin with stable structures. CA15, which exhibited a stronger inhibited effect on laryngeal cancer cells HEp-2 but a lower toxicity on hepatic cells HL-7702 in MTT assay, was selected for further analysis. The effects of CA15 on cell viability, proliferation, migration, apoptosis, and NF-κB activation were measured using MTT, Transwell migration, flow cytometry, Western blot, and immunofluorescence assays in HEp-2 cells. An NF-κB inhibitor, BMS-345541, as well as curcumin was also tested. Results showed that CA15 induced decreased toxicity towards HL-7702 cells compared to curcumin and BMS-345541. However, similar to BMS-345541 and curcumin, CA15 not only significantly inhibited proliferation and migration and induced caspase-3-dependent apoptosis but also attenuated TNF-α-induced NF-κB activation in HEp-2 cells. These results demonstrated that curcumin analogue CA15 exhibited anticancer effects on laryngeal cancer cells via targeting of NF-κB.

1. Introduction

Laryngeal carcinoma remains one of the most common malignancies of human beings. Laryngeal squamous cell carcinoma (LSCC) comprises more than 95% of laryngeal cancers whose morbidity reaches 4.1/100,000 worldwide [1, 2]. Chemotherapy is considered as the most effective treatment for LSCC besides surgery, and cis-platinum which prominently improves the survival rate of LSCC patients is widely used [3, 4]. However, many side effects of chemotherapy have been reported, including leukopenia and kidney failure [5]. Exploitation of new drugs showing greater therapeutic efficacy but relatively low toxicity is of great interest nowadays.

Nuclear factor of κB (NF-κB) is a transcription factor which can bind to specific sequence known as κB site [6]. It remains inactive by the inhibitor of κB (IκB) in the cytosol and could be activated by various carcinogenic factors including tumor necrosis factor-α (TNF-α) in cancer cells. The activation of myeloid differentiation protein 2 (MD2), a surface receptor of this signal, can induce the phosphorylation of IκB kinase (IKK). When IκB becomes phosphorylated, it dissociates from NF-κB, thus enabling NF-κB to translocate to the nucleus, bind to the κB site, and activate genes to prompt cancer development afterwards [7]. NF-κB is a potential target for cancer therapy.

Curcumin (diferuloylmethane), the major component isolated from the rhizomes of *Curcuma longa*, works as an anticancer agent in current researches [8]. Curcumin has shown impressive toxicity to LSCC in vitro and its inhibited effect on transcription factors NF-κB may contribute at least partly to its anticancer effect [9–12]. However, its low stability and poor bioavailability in vivo motivate the modification of the molecule to acquire improved potency

and physical properties [13]. It has been demonstrated that the β-diketone moiety probably results in the instability and rapid metabolism of curcumin [14].

Therefore, the monocarbonyl analogues of the curcumin (MCACs) which acquire a preferable stability were designed and synthesized in the present study. We carried out docking simulation of them with MD2 protein and screened out a promising compound named CA15 which exhibited a more notable effect against laryngeal cancer cells but a lower toxicity towards normal hepatic cells in comparison with curcumin. Furthermore, we investigated its anticancer effects against HEp-2 cells in detail and demonstrated that the agent could suppress NF-κB signal by inhibiting phosphorylation of IKK in HEp-2 cells.

2. Materials and Methods

2.1. Cell Culture. Human laryngeal squamous cell carcinoma HEp-2 cells and human hepatic cells HL-7702 were obtained from the Cell Bank of the Chinese Academy of Sciences (Shanghai, China). They were cultured in RPMI-1640 (Gibco) medium with 10% fetal bovine serum (FBS, Gibco) and 1% penicillin/streptomycin in a humidified atmosphere at 37°C with 5% CO_2. The medium was replaced every other day and cells in the following experiments were collected at the logarithmic growth phase.

2.2. Designs and Synthesis of MCACs. Chemical constructions of curcumin and MCACs were depicted in Figure 1(a). A total of 34 MCACs were designed and synthesized by Chemical Biology Research Center, School of Pharmaceutical Sciences, Wenzhou Medical University, Wenzhou, Zhejiang, China. Supplementary Figure (in Supplementary Material available online at https://doi.org/10.1155/2017/4751260) shows their chemical constructions. All the compounds were dissolved in dimethyl sulfoxide (DMSO) and stored at −20°C before being diluted into final concentration by the culture medium in the following experiments.

2.3. Docking of MCACs to the MD2 Structural Model. The crystal structure of the MD2 was retrieved from Protein Data Bank (PDB ID: 2E56). SYBYL X-2.0 software was used for the preparation of protein and compounds. The target protein was prepared through extracting ligand structure, removing water molecular, adding charge in termini treatment, and adding hydrogen. Adding hydrogen, adding charge, Powell energy gradient method, Tripos force field, and Gasteiger-Hückel system were used to minimize the MCACs.

2.4. Cell Viability Assay. HEp-2 and HL-7702 cells were plated in a density of 5000 cells/well in 100 ul medium containing FBS and indicated test compounds were added. After 72 h incubation, the cell viability was evaluated by MTT assay and repeated at least three times. Briefly, each well was added with 20 μl MTT solutions and then incubated for 4 h at 37°C in the dark. After removing the medium, 150 μl DMSO was added to each well. The absorbance was measured using an ELISA plate reader at 490 nm. Half-maximal inhibitory concentrations (IC_{50}) were determined using Sigma Plot 9.0 software (Systat Software Inc., CA) using the 4-parameter logistic function standard curve analysis for dose response.

2.5. Colony Formation Assay. Approximately 500 HEp-2 cells were cultured in medium containing 10% FBS at a final volume of 1 ml. Then the indicated compounds were added to the medium which was replaced 24 h later. After incubating for 7 days, cells were washed with PBS and fixed with 4% paraformaldehyde. The colonies were stained with crystal violet (Beyotime; Beijing, China) and photographed by a camera (Panasonic, Japan).

2.6. Transwell Migration Assay. Transwell migration chambers (8 μm pore size; BD Biosciences, USA) were purchased for observing the chemotactic motility of cells. Firstly, each top chamber was filled with 200 μL serum-free medium containing cells, while the bottom chamber was filled with 600 μL RMPI-1640 medium containing 10% FBS. The indicated compounds were then added to both chambers at the same concentration. After a 24 h incubation in a humidified atmosphere, nonmigrated cells were erased by cotton swabs. The migrated cells were fixed with 4% paraformaldehyde and stained with crystal violet solution. Images were photographed using an inverted microscope (Nikon; Japan) and the cells in at least 3 random microscopic fields were counted.

2.7. Measurements of Apoptosis. Approximately 1 million HEp-2 cells were incubated with indicated compounds for 12 h. Subsequently, cells were washed with cold PBS, harvested in binding buffer, and successively incubated with 3 μl Annexin V-FITC and 1 μl of PI (BD Biosciences, San Jose, CA, USA) for 15 min at room temperature in darkness. Apoptosis was determined by a Flow Cytometer (BD Biosciences, USA). For flow cytometric dot plot, Annexin V staining was set as the horizontal axis and PI staining was set as the vertical axis. Early apoptosis cells in the lower right quadrant and late apoptosis or necrotic cells in the upper right quadrant of the flow cytometric dot plot were both calculated.

2.8. Western Blot Analysis. Cells that needed examination were lysed in RIPA buffer to extract the total cellular protein. Protein concentration was determined and 60 μg of each protein sample was boiled at 100°C in SDS sample buffer for 10 min, electrophoresed on 10% SDS/PAGE gels, and then transferred to polyvinylidene difluoride (PVDF) membranes. Following protein transfer, the membranes were blocked by 5% skimmed milk proteins for 90 min, followed by incubation at 4°C overnight with the primary antibodies. Membranes were incubated with correlate secondary antibody at room temperature for 1 h on the second day, and specific protein bands were detected with an enhanced chemiluminescence (ECL) assay kit (BD, USA). The primary antibodies used were as follows: Cle-PARP, Bcl-2, Bax, p-IKK, IKK, IκB-α, and GAPDH (Santa Cruz Biotechnology, USA) and Cle-caspase-3 (Cell Signaling Technology, USA).

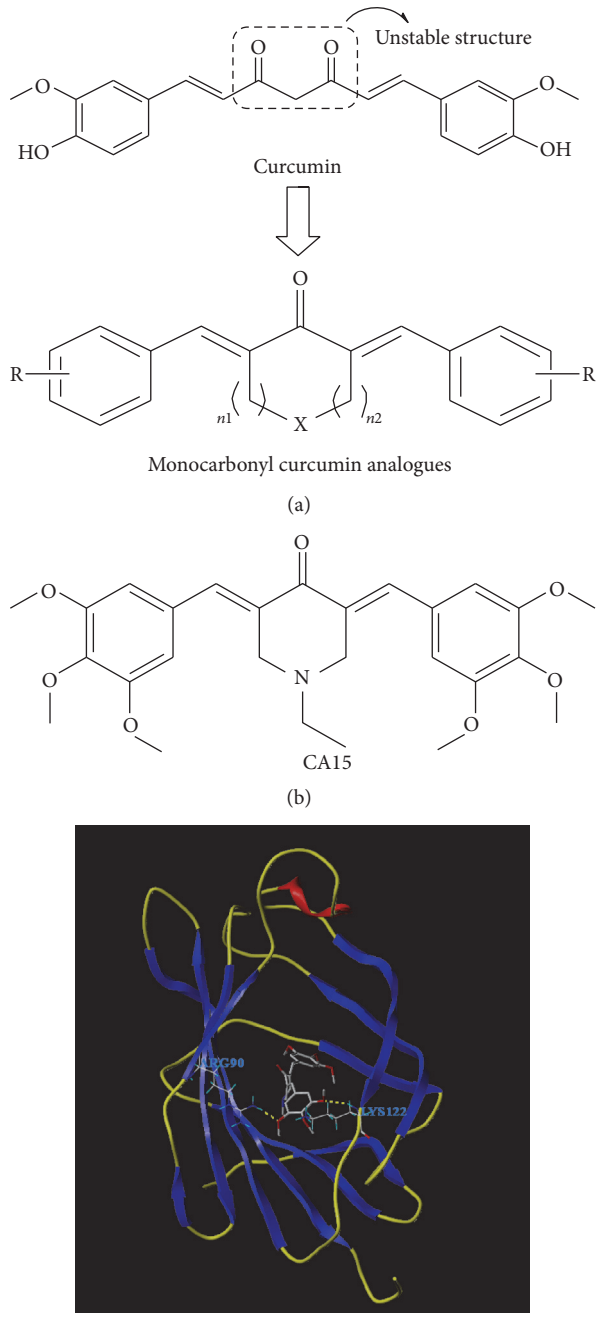


Figure 1: Design and synthesis of MCACs (a), chemical structure of curcumin analogue CA15 (b), and molecular docking of CA15 with MD2 (PDB ID 2E56). MCACs: monocarbonyl analogues of the curcumin.

2.9. Immunofluorescence. Approximately 500,000 HEp-2 cells were seeded on every glass slide in the 6-well plate and incubated with indicated compounds for 1 h. Cells were fixed in 4% paraformaldehyde for 15 min and permeated by 0.3% Triton X-100 (Beyotime; Beijing, China) for another 15 min. After blocking with 10% goat serum, slides were washed and incubated with anti-NF-κB-p65 antibody (Santa Cruz Biotechnology, USA) at 4°C overnight. Afterwards, they were incubated with PE-conjugated secondary antibody (Santa Cruz Biotechnology, USA) for 1 h at room temperature in the dark on the following day. After PBS washing, the slides were counterstained using 5 μg/ml DAPI solutions (Sigma, USA) for 5 min before being photographed by a fluorescent microscope (Nikon, Japan).

2.10. Statistical Analysis. All results are expressed as means ± standard errors from three independent experiments. The statistical analysis was performed by using Student's *t*-test or

Table 1: Docking of MCACs to the MD2 structural model.

Compounds	Acting sites	Total score
CA15	ARG^{90}, LYS^{122}	7.38
CA32	SER^{120}	6.40
CA33	SER^{120}	6.02
CA1	TYR^{102}	5.92
CA2	TYR^{102}	5.61
CA34	LYS^{122}	5.01
CA28	ARG^{90}	4.41

MCACs: monocarbonyl analogues of the curcumin; MD2: myeloid differentiation protein 2.

one-way analysis of variance (GraphPad Prism 5.0). $P < 0.05$ was considered to be statistically significant.

3. Results

3.1. CA15 Acted with MD2 Protein on Arg^{90} and Lys^{122} Residues and Achieved the Highest Total Score. The molecular docking experiments were performed at PH 7.0. As shown in Table 1, only 7 of these compounds had acting sites with MD2 protein using a molecular simulation. CA15 had the highest total score among 7 agents and might be the most probable candidate as an anticancer agent of NF-κB. Figure 1(b) depicted its chemical structure, and the computer-assisted simulation indicated that Arg^{90} and Lys^{122}, two amino residues of MD2 protein, were most likely to form hydrogen bonds with CA15 (Figure 1(c)).

3.2. CA15 Exhibited a More Preferable Cytotoxicity to HEp-2 Cells but a Lower Toxicity to HL-7702 Cells. The inhibitory rates of MCACs on HEp-2 and HL-7702 cells at the concentration of 20 μM were shown in Supplementary Table. CA15 was selected for further analysis, which expressed an inhibitory rate of 83.38 ± 4.79% on HEp-2 cells while only 36.51 ± 3.21% on HL-7702 cells at the concentration of 20 μM. Furthermore, IC_{50} values of CA15 were also tested by MTT. A highly selective inhibitor of NF-κB, BMS-345541 (BMS, Sigma) [15], was used as control as well as curcumin and its well-known analogues (EF24 and B19). Though without a significant difference, CA15 exhibited a more preferable cytotoxicity to HEp-2 cells but a lower toxicity to normal cells HL-7702 when compared to curcumin (Table 2). Stronger toxic effect on HL-7702 cells was observed in BMS compared to curcumin, which had the lowest IC_{50} value on HEp-2 cells (Table 2).

3.3. CA15 Inhibited Proliferation and Migration of HEp-2 Cells. To investigate the anticancer effects of CA15 on HEp-2 cells in detail, its inhibitory effect on colony formation of HEp-2 cells was examined firstly. As shown in Figure 2(a), CA15 suppressed colony formations of HEp-2 cells in a dose-dependent manner. By Transwell assay, we found that HEp-2 cells treated with curcumin or BMS showed obviously inhibited migration capability. Similarly, CA15 inhibited transmigration of HEp-2 significantly at the concentration of 20 μM ($P < 0.01$, Figure 2(b)).

3.4. CA15 Induced Apoptosis via Bax/Bcl-2 and Caspase-3-Dependent Pathway in HEp-2 Cells. We observed that CA15 could induce apoptosis in HEp-2 cells using Annexin V/PI. Apoptotic rates of HEp-2 cells treated with 5 μM CA15, 10 μM CA15, and 20 μM CA15 were 9.78 ± 0.93%, 26.18 ± 3.72%, and 31.53 ± 4.76%, respectively, which were all significantly increased compared to the control group (Figure 3(a)). Western blot analysis revealed that CA15 decreased the anti-apoptotic Bcl-2 protein, while it increased the proapoptotic Bax in a dose-dependent manner (Figure 3(b)). As expected, elevated Bax/Bcl-2 ratio led to cleavage and activation of PARP and caspase-3 (two markers for cell apoptosis) in HEp-2 cells (Figure 3(b)). Collectively, these results suggested that CA15 resulted in apoptosis induction via Bax/Bcl-2 and caspase-3-dependent pathway.

3.5. CA15 Inhibited TNF-α-Induced NF-κB Activation in HEp-2 Cells. Finally, whether CA15 could inhibit NF-κB activation in HEp-2 cells was investigated. As shown in Figure 4(a), TNF-α-induced phosphorylation of IKK was strikingly decreased in a dose-dependent manner following pretreatment with CA15. Phosphorylation of IKK leads to degradation of IκB. As expected, pretreatment with CA15 markedly reversed TNF-α-induced IκB degradation in HEp-2 cells in a dose-dependent manner (Figure 4(b)). NF-κB-p65 is able to translocate to the nucleus and raise its DNA-binding capacity without inhibition of IκB. Consequently, TNF-α vastly prompted the nuclear translocation of NF-κB-p65, whereas CA15 suppressed this process significantly apart from curcumin and BMS (Figure 4(c)).

4. Discussion

Although treatments of LSCC with surgery and chemoradiotherapy have been improved vastly in recent years [16, 17], the unintentional injuries on laryngeal functions and quality of life remain inevitable. As a natural compound isolated from plants, curcumin has a low toxicity but a remarkable anticancer effect on various malignancies such as lung cancer, liver cancer, and colon cancer [18–20]. Apart from that, curcumin is considered to be a repressor of LSCC growth via topical application [9]. However, the usage of curcumin is limited by the poor bioavailability due to its β-diketonate structure [13]. Although novel biocompatible nanosystems for curcumin delivery have been developed, the anticancer effects of curcumin nanoparticles could not last for a long time [21, 22]. The β-diketonate structure in curcumin is modified in MCACs (Figure 1(a)), which could improve its chemical stability in a different way [23].

Therefore, we synthesized 34 MCACs and screened out CA15 which performed a stronger inhibitory effect on laryngeal cancer cells HEp-2 but a lower toxicity on hepatic cells HL-7702 compared to curcumin (Table 2). A number of MCACs have been reported to have a similar biological effect

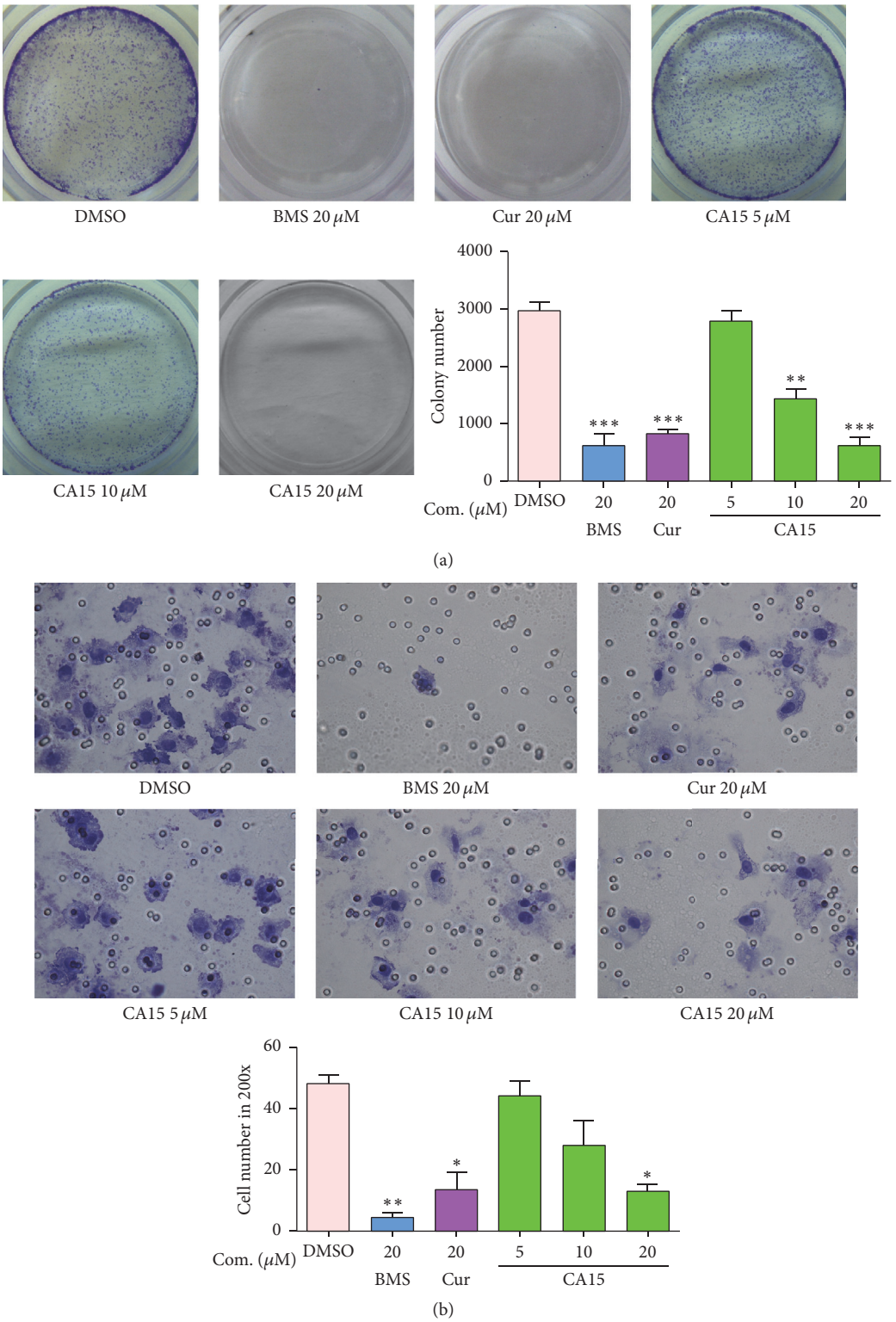


Figure 2: CA15 suppressed proliferation and migration in HEp-2 cells. (a) HEp-2 cells were treated with different concentrations (0–20 μM) of CA15, 20 μM Cur, and 20 μM BMS for 24 h. Cell proliferation was evaluated with colony formation assay. (b) HEp-2 cells were treated with different concentrations (0–20 μM) of CA15, 20 μM Cur, and 20 μM BMS for 24 h. Cell migration was examined using Transwell migration chambers and then observed under a microscope (magnification ×200). The graph displays means ± SEM of 3 independent experiments. $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$ versus DMSO group. Cur: curcumin; BMS: BMS-345541.

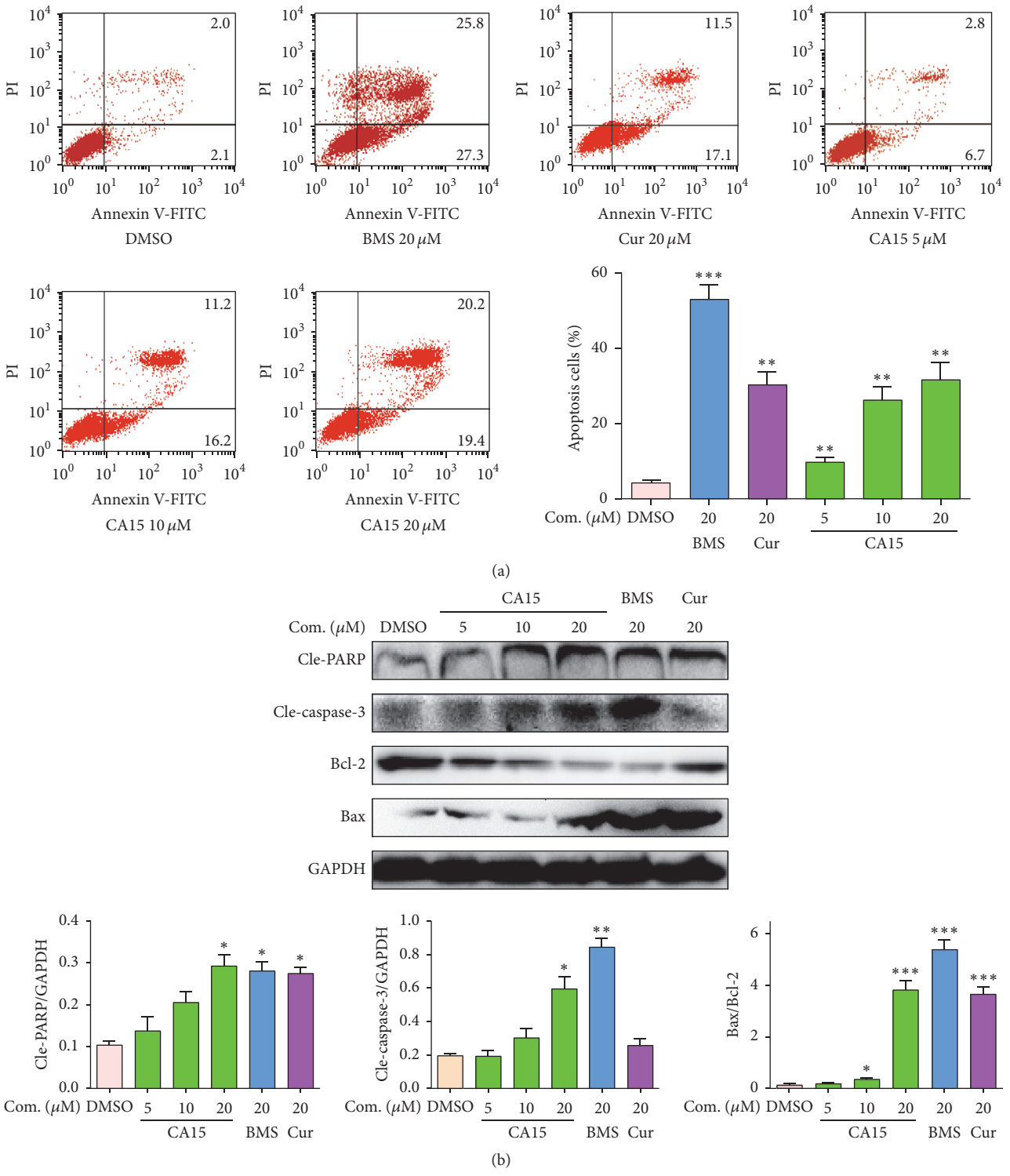


Figure 3: CA15 induced apoptosis by upregulating Bax/Bcl-2 ratio followed by caspase-3 cleavage in HEp-2 cells. HEp-2 cells were treated with different concentrations (0–20 μM) of CA15, 20 μM Cur, and 20 μM BMS for 24 h and analyzed by flow cytometry (a) or for 12 h and detected by Western blot (b). The graphs display means ± SEM of 3 independent experiments. $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$ versus DMSO group. Cur: curcumin; BMS: BMS-345541.

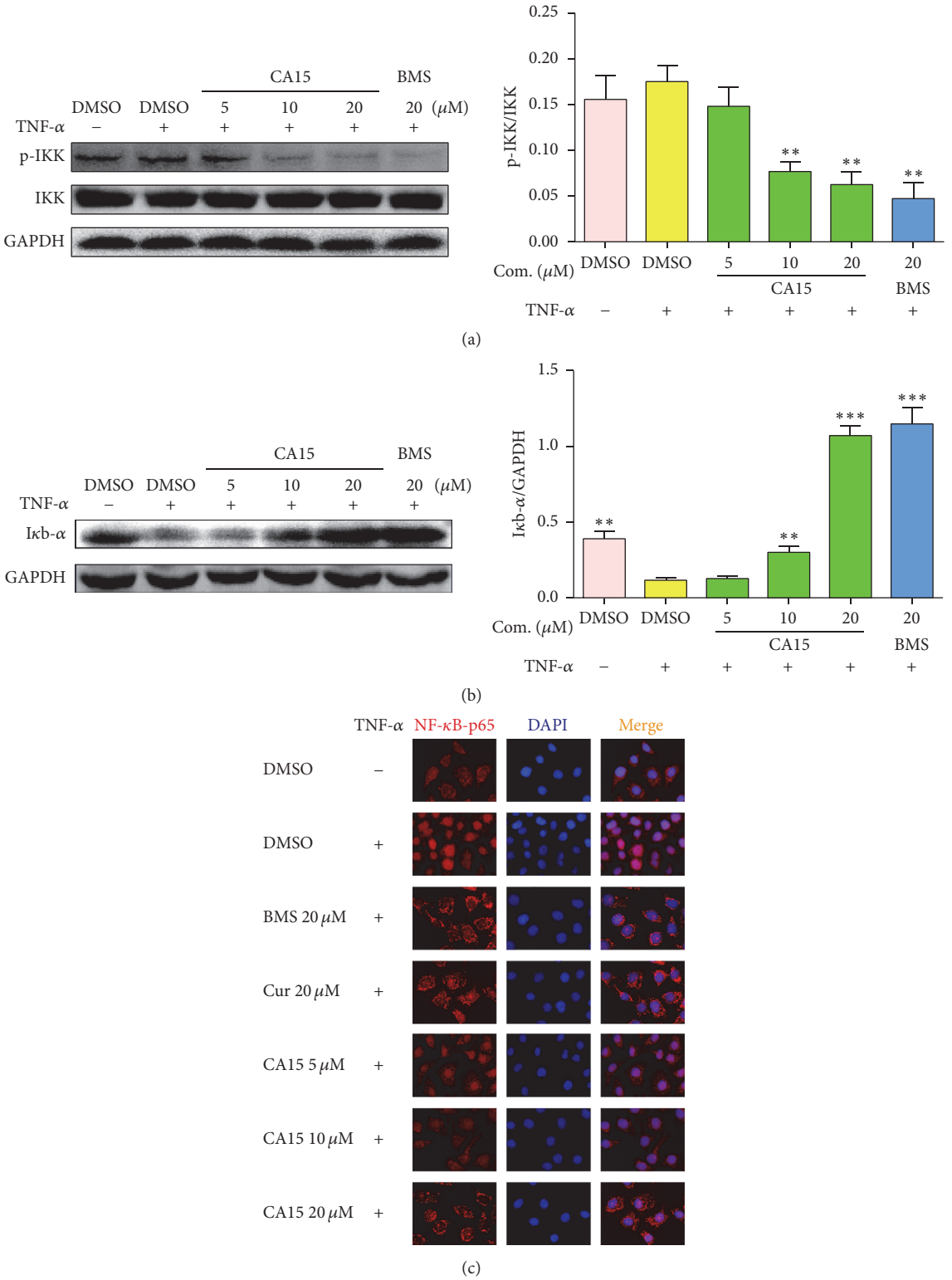


FIGURE 4: CA15 inhibited TNF-α-induced NF-κB activation in HEp-2 cells. (a and b) HEp-2 cells were pretreated with different concentrations (0–20 μM) of CA15 and 20 μM BMS for 1 h followed by incubation with TNF-α (5 ng/ml) for 15 min (a) or 30 min (b). The protein levels of p-IKK, IKK, and IκB were examined by Western blot. The graphs display means ± SEM of 3 independent experiments. $^{**}P < 0.01$ and $^{***}P < 0.001$ versus TNF-α treated group. (c) HEp-2 cells were pretreated with 20 μM CA15 and 20 μM BMS for 1 h followed by incubation with TNF-α (5 ng/ml) for 60 min. Cells were then incubated with NF-κB-p65 antibody and PE-conjugated secondary antibody (red), and the nuclei were stained with DAPI (blue). The photographs were obtained by fluorescence microscope (magnification ×200). TNF-α: tumor necrosis factor-α; Cur: curcumin; BMS: BMS-345541.

Table 2: The IC_{50} values of compounds towards HEp-2 and HL-7702 cells.

Group	HEp-2 (μM)	HL-7702 (μM)
Curcumin	21.37 ± 3.25	32.33 ± 6.21
CA15	16.59 ± 2.43	46.79 ± 8.77
EF24	14.09 ± 5.32	12.87 ± 2.73*
B19	11.69 ± 2.66	15.19 ± 2.76
BMS-345541	7.58 ± 0.41*	23.63 ± 3.24

HEp-2 and HL-7702 cells were treated with different concentrations (0–60 μM) of compounds for 72 h and then tested with MTT assay. Data are presented as means ± SEM of 3 independent experiments. * $P < 0.05$ versus curcumin group.

but a more stable chemical structure compared to curcumin [23–26]. As a curcumin analogue with greater bioavailability and biological activity, EF24 is able to inhibit proliferation of colon cancer cells and migratory of melanoma cells [27, 28]. Another curcumin analogue, B19, is also found to exert antiangiogenic activity and induce apoptosis of ovarian cancer cells [29, 30]. However, their toxicities towards normal cells are always neglected. They were both highly toxic to normal hepatic cells in our study. By contrast, CA15 treated HL-7702 cells that presented higher IC_{50} value, which indicated that CA15 might be a less toxic drug compared to curcumin (Table 2). Nevertheless, CA15, just like curcumin, was able to trigger caspase-3-dependent apoptosis in HEp-2 cells and exhibit remarkable inhibition on cell proliferation and migration.

As we know, whether a cell goes into the programmed cell death partly depends on the balance between proteins that promote cell viability (e.g., Bcl-2) and proteins that mediate apoptosis (e.g., PARP and Bax). Downregulation of Bcl-2 upregulates Bax and increases the Bax/Bcl-2 ratio, thereby inducing cleavage of PARP and caspase-3 and ultimately promoting hydrolysis of cytoskeletal proteins and nucleic acids [31]. NF-κB is a key transcription factor that targets Bcl-2 and regulates caspase-3-dependent apoptosis [32]. It has been reported that NF-κB activated cells exhibited enhanced expression of Bcl-2 protein, while NF-κB inhibition impaired Bcl-2 expression [33]. The dysregulation or chronic activation of NF-κB signaling restrains the programmed cell death of cancer cells, thus contributing a lot to the occurrence of many cancers including head and neck squamous cell carcinoma (HNSCC). A study reported that over 1,000 NF-κB targeting genes are differentially expressed in head and neck squamous cell carcinoma (HNSCC) in comparison with nonmalignant keratinocytes, along with the aberrant activation of NF-κB signaling [34]. These known genes are involved in the process of proliferation, diversification, angiogenesis, adhesion, and epithelial-mesenchymal transition of cancer cells, thus assisting tumors in evasion of apoptosis, sustained growth, invasion, and metastasis [34]. A number of factors may contribute to the activation of NF-κB in HNSCC, such as stimuli of tobacco and alcohol and infections of EB and HPV [35–37]. Activation of NF-κB prompts the invasion and metastasis process of HNSCC [38, 39] and is probably associated with poor outcome among these patients [39–41]. Accordantly, the increase of NF-κB associated cytokines, such as IL-6 and VEGF, may predict a poor prognosis in HNSCC patients [42–44]. Furthermore, cancer cells with defective NF-κB signaling seemed to be more sensitive to chemotherapy, which suggests that NF-κB should be probably involved in drug-resistance of HNSCC [45]. Taken together, NF-κB may promote cancer occurrence, progression, and drug resistance in HNSCC.

Currently, therapeutics based on targeting NF-κB are of considerable interest. Our study also demonstrated that CA15 could inhibit NF-κB signaling in HEp-2 cells. The mechanism of CA15 targeting NF-κB is like BMS, which acts as a highly selective inhibitor of IKK which binds at an allosteric site of the enzyme [15]. CA15 can inhibit the phosphorylation of IKK and hinder the degradation of IκB which binds to NF-κB-p65, thereby preventing NF-κB-p65 from translocating to the nucleus and regulating related genes (Figure 4). Several experiments demonstrated that intervention of NF-κB signal is a potential approach to cure LSCC [46, 47]. Agents such as PDTC, celecoxib, guggulsterone, and bortezomib have already showed impressive inhibitory effects on HNSCC via targeting NF-κB [48–51]. NF-κB might be a major focus of therapeutic intervention for LSCC treatments.

However, there is still limited knowledge regarding the specific mechanisms of CA15 against cancers. Also, its toxicity in vivo was not tested in our study. Further researches into the mechanisms and toxicity in vivo are warranted.

5. Conclusion

The current study reveals that CA15, a novel monocarbonyl curcumin analogue, exhibits preferable anticancer effects via targeting NF-κB but little toxicity to normal cells. NF-κB may be a potential target for LSCC treatment, and CA15, perhaps, has a potential therapeutic use in the treatment of laryngeal cancer in the future.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Jian Chen and Linlin Zhang contributed equally to this work.

Acknowledgments

This work was supported by National Natural Science Foundation of China (81272462 and 81402839) and Zhejiang Province Natural Science Funding of China (LY17H160059).

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA: Cancer Journal for Clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[2] J. S. Cooper, K. Porter, K. Mallin et al., "National cancer database report on cancer of the head and neck: 10-Year update," *Head and Neck*, vol. 31, no. 6, pp. 748–758, 2009.

[3] J. P. Pignon, J. Bourhis, C. Domenge, and L. Designe, "Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. MACH-NC Collaborative Group. Meta-Analysis of Chemotherapy on Head and Neck Cancer," *The Lancet*, vol. 355, pp. 949–955, 2000.

[4] A. A. Forastiere, H. Goepfert, M. Maor et al., "Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer," *The New England Journal of Medicine*, vol. 349, no. 22, pp. 2091–2098, 2003.

[5] S. Billan, O. Kaidar-Person, F. Atrash et al., "Toxicity of induction chemotherapy with docetaxel, cisplatin and 5-fluorouracil for advanced head and neck cancer," *The Israel Medical Association Journal*, vol. 15, no. 5, pp. 231–235, 2013.

[6] A. Gupta, R. Kumar, V. Sahu et al., "NFκB-p50 as a blood based protein marker for early diagnosis and prognosis of head and neck squamous cell carcinoma," *Biochemical and Biophysical Research Communications*, vol. 467, no. 2, pp. 248–253, 2015.

[7] B. Hoesel and J. A. Schmid, "The complexity of NF-κB signaling in inflammation and cancer," *Molecular Cancer*, vol. 12, no. 1, article 86, 2013.

[8] M. A. Adahoun, M. H. Al-Akhras, M. S. Jaafar, and M. Bououdina, "Enhanced anti-cancer and antimicrobial activities of curcumin nanoparticles," *Artificial Cells, Nanomedicine, and Biotechnology*, vol. 45, no. 1, pp. 98–107, 2016.

[9] A. Hu, J. J. Huang, X. J. Jin et al., "Curcumin suppresses invasiveness and vasculogenic mimicry of squamous cell carcinoma of the larynx through the inhibition of JAK-2/STAT-3 signaling pathway," *American Journal of Cancer Research*, vol. 5, pp. 278–288, 2015.

[10] H. Zhang, T. Yu, L. Wen, H. Wang, D. Fei, and C. Jin, "Curcumin enhances the effectiveness of cisplatin by suppressing CD133+ cancer stem cells in laryngeal carcinoma treatment," *Experimental and Therapeutic Medicine*, vol. 6, no. 5, pp. 1317–1321, 2013.

[11] Ö. Berrak, Y. Akkoç, E. D. Arisan, A. Çoker-Gürkan, P. Obakan-Yerlikaya, and N. Palavan-Ünsal, "The inhibition of PI3K and NFκB promoted curcumin-induced cell cycle arrest at G2/M via altering polyamine metabolism in Bcl-2 overexpressing MCF-7 breast cancer cells," *Biomedicine and Pharmacotherapy*, vol. 77, pp. 150–160, 2016.

[12] V. M. Duarte, E. Han, M. S. Veena et al., "Curcumin enhances the effect of cisplatin in suppression of head and neck squamous cell carcinoma via inhibition of IKKβ protein of the NFβB pathway," *Molecular Cancer Therapeutics*, vol. 9, no. 10, pp. 2665–2675, 2010.

[13] S. Manohar, S. I. Khan, S. K. Kandi et al., "Synthesis, antimalarial activity and cytotoxic potential of new monocarbonyl analogues of curcumin," *Bioorganic and Medicinal Chemistry Letters*, vol. 23, no. 1, pp. 112–116, 2013.

[14] M. J. C. Rosemond, L. St. John-Williams, T. Yamaguchi, T. Fujishita, and J. S. Walsh, "Enzymology of a carbonyl reduction clearance pathway for the HIV integrase inhibitor, S-1360: role of human liver cytosolic aldo-keto reductases," *Chemico-Biological Interactions*, vol. 147, no. 2, pp. 129–139, 2004.

[15] J. R. Burke, M. A. Pattoli, K. R. Gregor et al., "BMS-345541 is a highly selective inhibitor of IκB kinase that binds at an allosteric site of the enzyme and blocks NF-κB-dependent transcription in mice," *The Journal of Biological Chemistry*, vol. 278, no. 3, pp. 1450–1456, 2003.

[16] C. P. McMullen and R. V. Smith, "Treatment/comparative therapeutics: cancer of the larynx and hypopharynx," *Surgical Oncology Clinics of North America*, vol. 24, no. 3, pp. 521–545, 2015.

[17] N. Denaro, E. G. Russi, J. L. Lefebvre, and M. C. Merlano, "A systematic review of current and emerging approaches in the field of larynx preservation," *Radiotherapy and Oncology*, vol. 110, no. 1, pp. 16–24, 2014.

[18] M. Ye, J. Zhang, J. Zhang, Q. Miao, L. Yao, and J. Zhang, "Curcumin promotes apoptosis by activating the p53-miR-192-5p/215-XIAP pathway in non-small cell lung cancer," *Cancer Letters*, vol. 357, no. 1, pp. 196–205, 2015.

[19] C. Rana, H. Piplani, V. Vaish, B. Nehru, and S. N. Sanyal, "Downregulation of PI3-K/Akt/PTEN pathway and activation of mitochondrial intrinsic apoptosis by Diclofenac and Curcumin in colon cancer," *Molecular and Cellular Biochemistry*, vol. 402, no. 1-2, pp. 225–241, 2015.

[20] J. U. Marquardt, L. Gomez-Quiroz, L. O. A. Camacho et al., "Curcumin effectively inhibits oncogenic NF-κB signaling and restrains stemness features in liver cancer," *Journal of Hepatology*, vol. 63, no. 3, pp. 661–669, 2015.

[21] S. M. Masloub, M. H. Elmalahy, D. Sabry, W. S. Mohamed, and S. H. Ahmed, "Comparative evaluation of PLGA nanoparticle delivery system for 5-fluorouracil and curcumin on squamous cell carcinoma," *Archives of Oral Biology*, vol. 64, pp. 1–10, 2016.

[22] J.-J. Lee, S. Y. Lee, J.-H. Park, D.-D. Kim, and H.-J. Cho, "Cholesterol-modified poly(lactide-co-glycolide) nanoparticles for tumor-targeted drug delivery," *International Journal of Pharmaceutics*, vol. 509, no. 1-2, pp. 483–491, 2016.

[23] C. Zhao, Z. Liu, and G. Liang, "Promising curcumin-based drug design: mono-carbonyl analogues of curcumin (MACs)," *Current Pharmaceutical Design*, vol. 19, no. 11, pp. 2114–2135, 2013.

[24] Q. Weng, L. Fu, G. Chen et al., "Design, synthesis, and anticancer evaluation of long-chain alkoxylated mono-carbonyl analogues of curcumin," *European Journal of Medicinal Chemistry*, vol. 103, pp. 44–55, 2015.

[25] S. Sharma, M. K. Gupta, A. K. Saxena, and P. M. S. Bedi, "Triazole linked mono carbonyl curcumin-isatin bifunctional hybrids as novel anti tubulin agents: design, synthesis, biological evaluation and molecular modeling studies," *Bioorganic & Medicinal Chemistry*, vol. 23, no. 22, pp. 7165–7180, 2015.

[26] Z. Liu, L. Tang, P. Zou et al., "Synthesis and biological evaluation of allylated and prenylated mono-carbonyl analogs of curcumin as anti-inflammatory agents," *European Journal of Medicinal Chemistry*, vol. 74, pp. 671–682, 2014.

[27] A. L. Kasinski, Y. Du, S. L. Thomas et al., "Inhibition of IκB kinase-nuclear factor-κB signaling pathway by 3,5-bis(2-flurobenzylidene)piperidin-4-one (EF24), a novel monoketone analog of curcumin?" *Molecular Pharmacology*, vol. 74, no. 3, pp. 654–661, 2008.

[28] P. Zhang, H. Bai, G. Liu et al., "MicroRNA-33b, upregulated by EF24, a curcumin analog, suppresses the epithelial-to-mesenchymal transition (EMT) and migratory potential of melanoma cells by targeting HMGA2," *Toxicology Letters*, vol. 234, no. 3, pp. 151–161, 2015.

[29] W. Qu, J. Xiao, H. Zhang et al., "B19, a novel monocarbonyl analogue of curcumin, induces human ovarian cancer cell apoptosis via activation of endoplasmic reticulum stress and the autophagy signaling pathway," *International Journal of Biological Sciences*, vol. 9, no. 8, pp. 766–777, 2013.

[30] L. Sun, J. Liu, S.-S. Lin et al., "Potent anti-angiogenic activity of B19—a mono-carbonyl analogue of curcumin," *Chinese Journal of Natural Medicines*, vol. 12, no. 1, pp. 8–14, 2014.

[31] A. K. Alam, A. S. Hossain, M. A. Khan et al., "The antioxidative fraction of white mulberry induces apoptosis through regulation of p53 and NFκB in EAC cells," *PLoS ONE*, vol. 11, no. 12, Article ID e0167536, 2016.

[32] C. V. Alexander-Savino, M. S. Hayden, C. Richardson, J. Zhao, and B. Poligone, "Doxycycline is an NF-kappa B inhibitor that induces apoptotic cell death in malignant T-cells," *Oncotarget*, vol. 7, pp. 75954–75967, 2016.

[33] Q. L. Ge, S. H. Liu, Z. H. Ai et al., "RelB/NF-κB links cell cycle transition and apoptosis to endometrioid adenocarcinoma tumorigenesis," *Cell Death and Disease*, vol. 7, article e2402, 2016.

[34] B. Yan, X. Yang, T.-L. Lee et al., "Genome-wide identification of novel expression signatures reveal distinct patterns and prevalence of binding motifs for p53, nuclear factor-κB and other signal transcription factors in head and neck squamous cell carcinoma," *Genome Biology*, vol. 8, article R78, 2007.

[35] C. T. Allen, J. L. Ricker, Z. Chen, and C. Van Waes, "Role of activated nuclear factor-κB in the pathogenesis and therapy of squamous cell carcinoma of the head and neck," *Head and Neck*, vol. 29, no. 10, pp. 959–971, 2007.

[36] Z. Chen, B. Yan, and C. Van Waes, "Role of the NF-κB transcriptome and proteome as biomarkers in human head and neck squamous cell carcinomas," *Biomarkers in Medicine*, vol. 2, no. 4, pp. 409–429, 2008.

[37] T. L. Lee, J. Yeh, J. Friedman et al., "A signal network involving coactivated NF-κB and STAT3 and altered p53 modulates BAX/BCL-XL expression and promotes cell survival of head and neck squamous cell carcinomas," *International Journal of Cancer*, vol. 122, no. 9, pp. 1987–1998, 2008.

[38] F. S. Giudice, A. M. da Costa Dal Vechio, A. C. Abrahão, F. F. Sperandio, and D. D. S. Pinto-Junior, "Different expression patterns of pAkt, NF-κB and cyclin D1 proteins during the invasion process of head and neck squamous cell carcinoma: an in vitro approach," *Journal of Oral Pathology and Medicine*, vol. 40, no. 5, pp. 405–411, 2011.

[39] Z.-J. Zhao, L. Peng, F.-Y. Liu, L. Sun, and C.-F. Sun, "PKCα take part in CCR7/NF-κB autocrine signaling loop in CCR7-positive squamous cell carcinoma of head and neck," *Molecular and Cellular Biochemistry*, vol. 357, no. 1-2, pp. 181–187, 2011.

[40] P. Balermpas, Y. Michel, J. Wagenblast et al., "Nuclear NF-κB expression correlates with outcome among patients with head and neck squamous cell carcinoma treated with primary chemoradiation therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 86, no. 4, pp. 785–790, 2013.

[41] Y. K. Mburu, A. M. Egloff, W. H. Walker et al., "Chemokine receptor 7 (CCR7) gene expression is regulated by NF-κB and activator protein 1 (AP1) in metastatic squamous cell carcinoma of head and neck (SCCHN)," *The Journal of Biological Chemistry*, vol. 287, no. 5, pp. 3581–3590, 2012.

[42] Z. Chen, P. S. Malhotra, G. R. Thomas et al., "Expression of proinflammatory and proangiogenic cytokines in patients with head and neck cancer," *Clinical Cancer Research*, vol. 5, no. 6, pp. 1369–1379, 1999.

[43] C. Allen, S. Duffy, T. Teknos et al., "Nuclear factor-κB-related serum factors as longitudinal biomarkers of response and survival in advanced oropharyngeal carcinoma," *Clinical Cancer Research*, vol. 13, no. 11, pp. 3182–3190, 2007.

[44] C. H. Druzgal, Z. Chen, N. T. Yeh et al., "A pilot study of longitudinal serum cytokine and angiogenesis factor levels as markers of therapeutic response and survival in patients with head and neck squamous cell carcinoma," *Head and Neck*, vol. 27, no. 9, pp. 771–784, 2005.

[45] N. Umemura, J. Zhu, Y. K. Mburu et al., "Defective NF-κB signaling in metastatic head and neck cancer cells leads to enhanced apoptosis by double-stranded RNA," *Cancer Research*, vol. 72, no. 1, pp. 45–55, 2012.

[46] Y.-L. Liu and Y.-N. Sun, "Down-regulation of testes-specific protease 50 induces apoptosis in human laryngocarcinoma HEp2 cells in a NF-κB-mediated pathway," *Molecular Biology Reports*, vol. 41, no. 12, pp. 7743–7747, 2014.

[47] H. Jin, W. Lou, and J. Sang, "Inhibitory effect of NF-kappaB p65siRNA on human laryngeal carcinoma xenograft model in nude mice," *Journal of Clinical Otorhinolaryngology, Head, and Neck Surgery*, vol. 27, no. 15, pp. 836–838, 2013.

[48] D. Santos Pinto, A. Abrahao, and A. Del Vecchio, "Expression analysis of proteins Cox-2, Akt, NF-kappa B, in head and neck squamous cell carcinoma (SCC) cells treated with Celecoxib," *Virchows Archiv*, vol. 459, p. S118, 2011.

[49] J. Gilbert, J. W. Lee, A. Argiris et al., "Phase II 2-arm trial of the proteasome inhibitor, PS-341 (bortezomib) in combination with irinotecan or PS-341 alone followed by the addition of irinotecan at time of progression in patients with locally recurrent or metastatic squamous cell carcinoma of the head and neck (E1304): a trial of the Eastern Cooperative Oncology Group," *Head and Neck*, vol. 35, no. 7, pp. 942–948, 2013.

[50] M. A. Macha, A. Matta, S. S. Chauhan, K. W. Michael Siu, and R. Ralhan, "Guggulsterone (GS) inhibits smokeless tobacco and nicotine-induced NF-κB and STAT3 pathways in head and neck cancer cells," *Carcinogenesis*, vol. 32, no. 3, pp. 368–380, 2011.

[51] M. Yan, Q. Xu, P. Zhang, X.-J. Zhou, Z.-Y. Zhang, and W.-T. Chen, "Correlation of NF-κB signal pathway with tumor metastasis of human head and neck squamous cell carcinoma," *BMC Cancer*, vol. 10, article 437, 2010.

Hindawi
BioMed Research International
Volume 2017, Article ID 3923865, 11 pages
https://doi.org/10.1155/2017/3923865

Research Article

Prediction of Radix Astragali Immunomodulatory Effect of CD80 Expression from Chromatograms by Quantitative Pattern-Activity Relationship

Michelle Chun-har Ng,[1] Tsui-yan Lau,[2] Kei Fan,[1] Qing-song Xu,[3] Josiah Poon,[4] Simon K. Poon,[4] Mary K. Lam,[5] Foo-tim Chau,[2] and Daniel Man-Yuen Sze[6]

[1]*Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
[2]*Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
[3]*School of Mathematics and Statistics, Central South University, Changsha 410083, China*
[4]*School of Information Technologies, The University of Sydney, Lidcombe, NSW, Australia*
[5]*Faculty of Health, University of Technology Sydney, Ultimo, NSW, Australia*
[6]*School of Health and Biomedical Sciences, RMIT University, Melbourne, VIC, Australia*

Correspondence should be addressed to Foo-tim Chau; foo-tim.chau@polyu.edu.hk
and Daniel Man-Yuen Sze; daniel.sze@rmit.edu.au

Received 9 September 2016; Revised 15 December 2016; Accepted 15 January 2017; Published 28 February 2017

Academic Editor: Adair Santos

The current use of a single chemical component as the representative quality control marker of herbal food supplement is inadequate. In this CD80-Quantitative-Pattern-Activity-Relationship (QPAR) study, we built a bioactivity predictive model that can be applicable for complex mixtures. Through integrating the chemical fingerprinting profiles of the immunomodulating herb *Radix Astragali* (RA) extracts, and their related biological data of immunological marker CD80 expression on dendritic cells, a chemometric model using the Elastic Net Partial Least Square (EN-PLS) algorithm was established. The EN-PLS algorithm increased the biological predictive capability with lower value of RMSEP (11.66) and higher values of R_p^2 (0.55) when compared to the standard PLS model. This CD80-QPAR platform provides a useful predictive model for unknown RA extract's bioactivities using the chemical fingerprint inputs. Furthermore, this bioactivity prediction platform facilitates identification of key bioactivity-related chemical components within complex mixtures for future drug discovery and understanding of the batch-to-batch consistency for quality clinical trials.

1. Introduction

A large pool of medicinal plants from Chinese herbal medicines (CHM) has a long historical clinical practice for more than 2000 years ago. However, the underlying mechanisms of action of the CHM remain largely unknown except the few examples of taxol [1] for anticancer, artesunate [2] for malaria treatment, and arsenic trioxide [3] for leukemia treatment. While these three herbal derived single compounds are responsible for the effective therapies, however, for most of the other clinically useful CHM, the mechanisms of action have been considered as that of "multicompound multitarget." The use of herbal formula by combining a few herbs based on the Chinese medicine theory further adds to this complexity. Thus, there exists a wide range of possible chemical compounds in each single herb or complex formula that may contribute to the clinical efficacy, but this crucial information is basically unknown at the moment. This lack of understanding of the active compounds and their targets in turn makes the quality control aspect of ensuring the batch-to-batch consistency of CHM difficult if not impossible.

Up to now, a CHM product PHY906 which is undergoing phase 2 clinical trial and being marketed as an adjuvant to chemotherapy attempted to address the batch-to-batch

consistency issue [4, 5]. The researchers established a platform of "Phytoceutica" to address the similarity index of products of different batches for both of their chemical fingerprinting using liquid chromatography-mass spectrometry (LCMS) and the biological fingerprinting by microarray profiling. While this platform suffices for the purpose of quality control, it is not powerful enough to help the identification of compounds that are actually related to the mechanism of actions, such as reducing chemotherapy-induced gastrointestinal toxicity in mice [4] and the clinically favorable outcomes in cancer [5]. Hence, there is an increasing attention in research development to evaluate the mixtures of compounds from the CHM extracts as a whole with the bioactivity for developing modern drugs.

The conventional component-based quality control approach may overestimate the therapeutic value of some highly representative components while the minor components are ignored for their active roles or masked in the crude extract [6]. Accordingly selecting only the major chemical components as the standard markers is not adequate to explain the total therapeutic effect of the CHM. More importantly, as different constituents may contribute to different therapeutic activities, therefore a single CHM may possess multiple therapeutic activities. Understanding the quantitative relationship between the multiple chemical constituents of a single CHM with the corresponding bioactivity is becoming imperative.

In the last few years, different international research teams have attempted to address this quality control of herbal medicines issue by both the chemical and biological fingerprinting approaches. For instance, in China, Yan and his colleagues [7] studied 28 samples of Radix Tinosporae for analgesic bioactivities on mice. Chen and his colleagues used 32 combinations of 5-herb CHM mixture including RA and studied antiplatelet in SD rats [8]. Another research group of Jiang is based on 31 batches of curcuma volatile oil to study antitumor in vitro effects [9]. In Belgium, Tistaert and his colleagues reported similar approach using 39 Mallotus extracts and examined the related cytotoxicity [10]. Also in Singapore, Ching and his colleagues used 6 different solvent systems to extract the A. *elliptica* leaves and studied the corresponding antiplatelet activities [11].

Our laboratory has also developed comprehensive methods with multicomponent quantification such as pattern-based approaches through chemometric data processing techniques that have been used for the identification of contributing elements within a mixture [12, 13]. In this study, we adopted and expanded our laboratory's chemometric methodology named Quantitative-Pattern-Activity-Relationship (QPAR) to study an immunomodulatory herbal medicines, Radix Astragali (RA, or commonly known as Huangqi) [12].

QPAR is a computer-assisted platform based on the application of statistics and data analytical methods for model development [12]. It simply colligates the extract of a single herb as chemical fingerprint with the corresponding biological activity. A statistical mathematical model is then built for revealing the valuable information of CHM related to the corresponding bioactivity. These developed models can also be used for predicting the biological activity of a HM based solely on its intact chemical fingerprint.

It is well known that RA is one of the most widely used CHM for the enhancement of "qi" based on Chinese medicine theory. About its related mechanisms of action, a few publications have demonstrated that RA is related to the increase of both humoral immunity [14] and cellular immunity in our body [15, 16] or immunomodulatory as a whole [17–20]. RA has also been shown to exert an anticarcinogenic effect in carcinogen-treated mice through activation of cytotoxic activity and the production of cytokines [21]. We have previously published that dendritic cells (DCs), as the most important professional antigen presenting cells in anticancer immunity, have been found to be defective in cancer patients [22, 23]. Furthermore, this defectiveness increased when cancer progressed to more advanced stage. It is known that CD80 is the most important costimulatory molecules on the surface of DCs to provide the crucial second signal for the proper stimulation of cancer antigen-specific naive T cells.

Therefore, in this project, we harness the knowledge of the prior QPAR methodology and the CD80 flow cytometric bioactivity platform. By producing more than 70 crude extracts of RA of varied components, we aimed to build a CD80-QPAR model of RA. With this model, we can demonstrate the model's predictability with the chromatogram alone of any new RA preparations as an input, and the corresponding bioactivity can be accurately predicted. This knowledge is important for the determination of the levels of bioactivity-related quality control chemical markers in herbal extracts to be used in clinical trials.

2. Materials and Methods

2.1. Preparation of Radix Astragali (RA) Extracts, Reagents, and Reference Compound. Three batches of raw RA (RA-A, RA-B, and RA-C) were used to prepare 72 extracts in total (24 extracts each) according to a modified extraction method based on the Chinese Pharmacopoeia. Briefly, 4 g raw herb was preimmersing with bidistilled water (100, 150, 200, and 250 mL) for 12 hr and refluxed for 0, 1, 2, 3, and 4 hrs. The mixtures were then filtered and concentrated under a rotary evaporator (Brand, Germany). RA extracts were finally obtained after lyophilisation. Each extract was stored under low humidity condition and was kept for biological assay within 3 months. All the extracts before chromatographic analysis and biological assay were filtered under 0.2 μm filter. Bidistilled water was produced in-house by Milli-Q® Advantage A10 water purification systems (Millipore; USA) and filtered with 0.22 μm Millipak®. All other chemicals and reagents used were of analytical grade unless indicated otherwise.

2.2. THP-1 Dendentic Cell (DC) Functional Flow Cytometric Platform. THP-1 was used as a convenient robust source of DC in this in vitro DC functionality flow cytometric study based on our previous method [24, 25]. Briefly, THP-1 cells were cultured in RPMI-1640 (Invitrogen, USA) supplemented with 10% foetal bovine serum (Gibco, USA) and

100 U/mL penicillin/streptomycin (Caisson, USA) at 37°C with 5% CO_2. A total of 3×10^5 THP-1 cells/well with 200 μL completed RPMI medium in 96-well flat-bottom plates were treated with 5 μL dried RA extracts in the final concentration of 1.5 mg/mL for 24 and 48 hrs. The untreated cell treated with DDI was used as a control, whereas Lipopolysaccharide (LPS) (Sigma, USA), a bacterial cell wall component, was used as a positive control. The treated cells were harvested and stained with fluorescence-conjugated monoclonal antibodies of specificity against CD80 (BD, USA) for 20 min at 4°C and propidium iodide (PI) staining for live cell discrimination. Data were then acquired on a FC500 Flow cytometry (Beckman Coulter, USA) and the results were analyzed using FlowJo software (USA) package. The percentage change of the effect of each RA extract resulted from the comparison of the untreated control, which was considered as 0%.

2.3. HPLC Instrumentation and Chromatographic Conditions. The HPLC system used for chemical fingerprinting consisted of an Agilent Series 1100 HPLC system (Agilent; USA) and Agilent series 6300 Ion Trap VL LC-DAD-MS instruments, with a Hypersil ODS column (250 mm × 4.6 mm, 5 μm) (Thermo Fisher Scientific; USA) and autosampler. The system was equipped with a HP1100 diode array detector. Chromatographic separation of the RA extracts was performed using a gradient elution based on a mobile phase consisting of (A) HPLC graded Acetonitrile (Tedia, USA) and (B) 0.1% acetic acid in bidistilled water. The gradient elution was carried out by varying mobile phase (A) from 0 to 10% (0–15 min), from 10 to 30% (15–30 min), followed by isocratic for 15 mins, then from 30 to 60% (45–60 min), and finally isocratic for 10 min. The mobile phase was pumped through the column at a flow rate of 0.8 mL min^{-1}. Analyses were performed at ambient temperature and detection wavelength was carried out at 200, 254, 270, 300, and 360 nm. The injection volume was 20 μL. Each extract was run three times in order to validate the repeatability and linearity.

2.4. QPAR Model Development and Statistical Analysis. The QPAR model development techniques were based on our published paper [12] and the workflow was illustrated in Supplementary Figure 1 (in Supplementary Material available online at https://doi.org/10.1155/2017/3923865). In brief, data of chemical fingerprint and immunomodulatory effect of the 72 RA extracts were individually collected. The chemical fingerprint of each extracts was preprocessed using "The Fingerprint Analysis Software" developed by the Research Centre of Modernization of Traditional Chinese Medicine of the Central South University, Changsha, China. The total extracts were divided into two sets based on Kennard and Stones algorithm [26], a training set embracing two-thirds of the total extracts (48 extracts) for QPAR model building and a test set consisting of the rest for model validation. Partial Least Square (PLS) methods were coded and executed in MATLAB for building up the QPAR predictive models.

3. Results

3.1. Scheme of CD80-QPAR Chemometrics Platform Development. The workflow of the model development in this study is presented in Supplementary Figure 1. With the connection of the known chemical and biological data from RA extracts, a model was then established. This model was used to predict the unknown biological activity of any new RA extract by simply providing the chemical fingerprint of that RA extract. The details of the data collection, model development and refinement, and the quantitative assessment are shown in the following.

3.2. Chemical Data Collection and Preprocessing. It has been observed that higher amount of potential active ingredients such as isoflavonoids and astragalosides can be extracted using reflux system compared with ultrasonication [27]. Therefore, using uniform design technique, the extraction factors were included reflux time and the solvent volume. By varying the two factors, a total of 72 extracts from 3 different batches of RA were prepared for this study. The combination of the reflux time and the solvent volumes was shown in Supplementary Table 1. The average extraction yield in percentage was 37.1 ± 4.2% of 4 g dry herb. The extracts were then run through HPLC and the components were showed as a chromatogram collected by using DAD (Detection range from 190 to 400 nm). This HPLC-DAD chromatogram was called chemical fingerprint (Supplementary Figure 2). Chemical fingerprint preprocessing of each extracts was essential for baseline correction and peaks alignment before QPAR data processing. This procedure was carried out by "The Fingerprint Analysis Software," as mentioned in the Method. Supplementary Figure 2 showed the HPLC-DAD chromatogram of all the RA extracts from the three batches before and after data preprocessing.

3.3. Similarity Analysis within Different RA Batches. To examine the variation of individual extracts prepared from different condition within the same batch, similarity analysis was employed to compare their chemical fingerprints. A median chemical fingerprint was computed as a reference fingerprint from each batch for this similarity analysis and three of them were shown in Figure 1. Each extract was then compared with the reference fingerprint of the same batch and the degree of the similarity was calculated quantitatively as similarity index or SI (%). The result showed that the SI value of each extract within the same batch was in the range of 88.3%–99.0%. The average SI value within extracts from three batches is 95.8 ± 3.0% (RA-A); 96.1 ± 2.3% (RA-B); and 95.8 ± 2.1% (RA-C), respectively (Supplementary Table 2 and Supplementary Figure 3). This result indicated the low variation of component difference between extracts from their respective batch. The extracts shared similar chromatographic patterns in comparison with their three groups, although they were obtained under different preparation conditions including refluxing time and solvent volume.

To compare the similarity between batches, the SI values of them were also calculated. Low variances were found between batches; batches B (99.9%) and C (98.8%) have

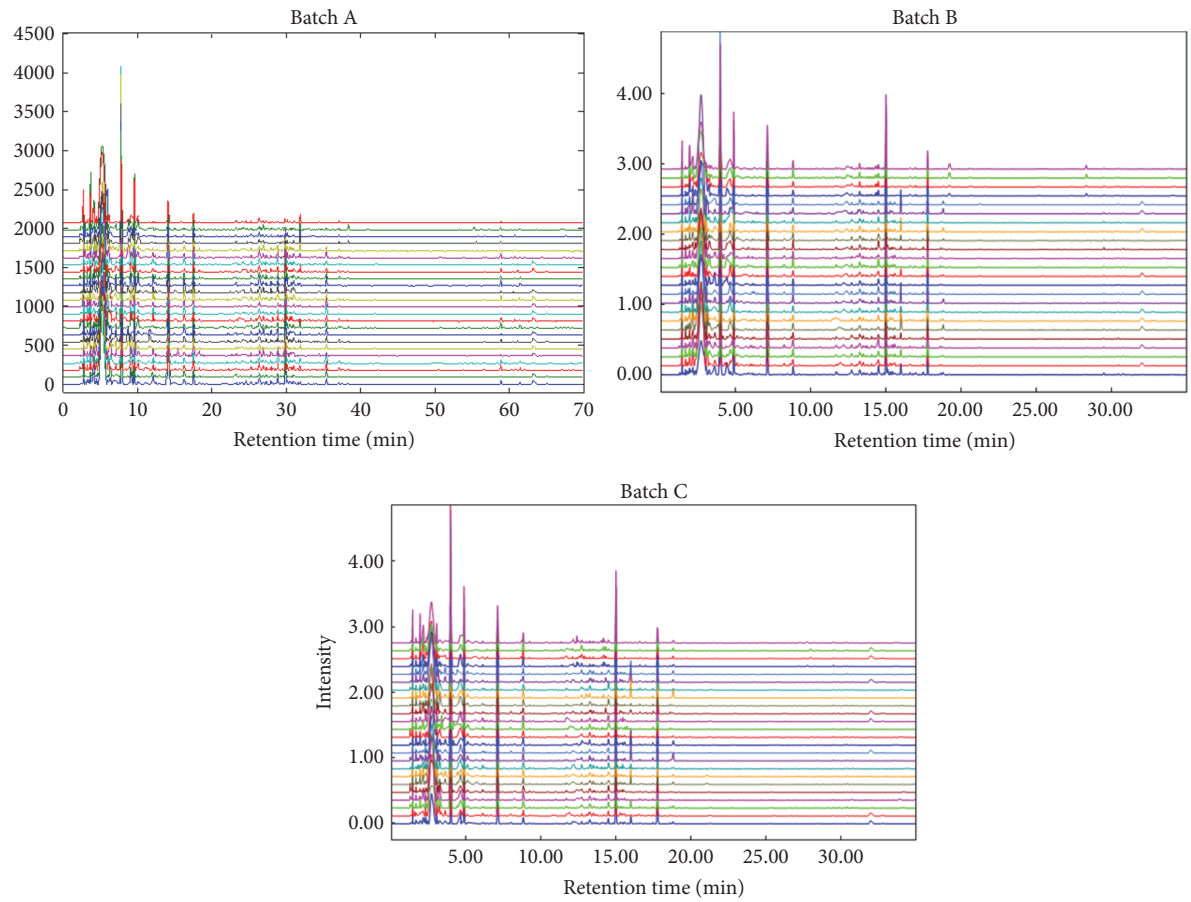


Figure 1: The HPLC-DAD chromatographic profiles of each RA extract from batches A, B, and C.

higher similarity on average than batch A (96.5%). This is not surprising, since batches B and C are from the same raw material; however, batch A is a stock from another source. This may explain the slight differences of chemical composition of batch A from that of batches B and C.

3.4. Biological Data Collection: Immunomodulatory Activity Represented by the Change of CD80 Expression Level. The biological activity in this study was the immunomodulatory effect of the RA extracts on THP-1 cell (Figure 2). It was showed as the expression level surface marker CD80. THP-1 is a human acute monocytic leukemia cell line and was used as a convenient robust dendritic cell (DC) platform for in vitro DC functionality flow cytometric study [25]. The immunomodulatory effect (relative change of CD80 expression to the blank, %, after standardization) of each RA extract from three batches on THP-1 cells were showed in Table 1. Interestingly, using the post hoc, LSD or Bonferroni analysis the biological activities were differences between three batches significantly, although similarity analysis indicated the similar chemical composition between batches (Supplementary Table 3).

3.5. Pivotal Role of Dendritic Cells in Regulation of Tumor-Specific Immune Responses by the Expression of the Costimulatory Surface Molecule CD80. The aqueous RA extracts were cocultured with THP-1 cells for 48 hours and the level of CD80 expression of the cells was detected by FACS analysis. The geometrical means (G means) of the relative fluorescence intensity indicated the CD80 expression level, and the normalized percentage change in CD80 expression from the treatment of various RA extracts was calculated by dividing the CD80 expression level of the treated assay with that of the one treated with double distilled water (DDI) (Figure 2). Lipopolysaccharide (LPS) was used to treat the cells as the positive controls. There was no activity found in the assay treated with DDI (0%), whereas the expression level of CD80 on THP-1 cells treating with the positive control, LPS, was upregulated to 51.2% ± 12.8. The ranges of the CD80 expression change in RA-A, RA-B, and RA-C are −14.7 to +30.7%, −19.3 to +20.6%, and −7.2 to +57.3%, respectively (Table 1).

Although the similarity analysis indicated the common pattern of chemical component in 72 RA extracts, the immunological activities were significantly different among three batches (Supplementary Table 3). The immunomodulatory effect of the RA extracts from batch C was significantly different from that of batch A ($p < 0.001$) and batch B ($p = 0$). The modulating effect of CD80 expression on THP-1 cells was also significantly different between batches A and B. This t-test analysis strongly indicated that the bioactivity capacities from batch C were significantly higher than that

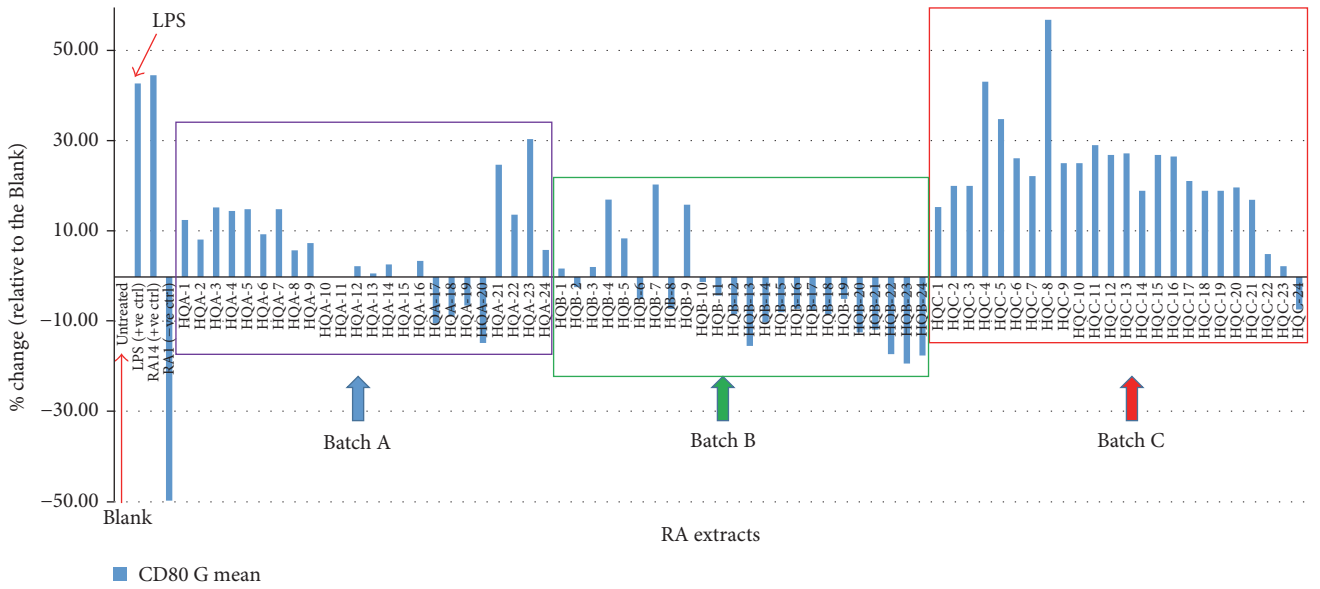


FIGURE 2: The immunomodulatory effect (relative percentage change of CD80 expression to the blank, after standardization) of each RA extract from three batches on THP-1 cell.

TABLE 1: The immunomodulatory effect (relative change of CD80 expression to the blank, %, after standardization) of each RA extract from three batches on THP-1 cell.

Batch A		Batch B		Batch C	
Sample	Act. activity (%)	Sample	Act. activity (%)	Sample	Act. activity (%)
A1	+12.70	B1	+1.87	C1	+15.58
A2	+8.33	B2	−2.25	C2	+20.29
A3	+15.48	B3	+2.25	C3	+20.29
A4	+14.68	B4	+17.23	C4	+43.48
A5	+15.08	B5	+8.61	C5	+35.14
A6	+9.52	B6	−4.87	C6	+26.45
A7	+15.08	B7	+20.60	C7	+22.46
A8	+5.95	B8	−7.12	C8	+57.25
A9	+7.54	B9	+16.10	C9	+25.36
A10	0	B10	−1.12	C10	+25.36
A11	0	B11	−4.12	C11	+29.35
A12	+2.38	B12	−8.24	C12	+27.17
A13	+0.79	B13	−15.36	C13	+27.54
A14	+2.78	B14	−10.11	C14	+19.20
A15	0	B15	−7.87	C15	+27.17
A16	+3.57	B16	−7.12	C16	+26.81
A17	−10.32	B17	−7.49	C17	+21.38
A18	−8.73	B18	−8.24	C18	+19.20
A19	−6.35	B19	−4.87	C19	+19.20
A20	−14.68	B20	−12.36	C20	+19.93
A21	+25.00	B21	−11.75	C21	+17.17
A22	+13.86	B22	−17.17	C22	+5.12
A23	+30.72	B23	−19.28	C23	+2.41
A24	+6.02	B24	−17.47	C24	−7.23
Avg +6.23		*Avg* −4.17		*Avg* +22.75	
SD 10.75		*SD* 10.78		*SD* 12.66	
Max +30.72		*Max* +20.60		*Max* +57.25	
Min −14.68		*Min* −19.28		*Min* −7.23	

Table 2: The results of the models built by three algorithms (PLS and EN-PLS).

Model	# of variables	Optimum # of PLS components	Training set			Test set	
			R_t^2	RMSET	RMSECV	R_p^2	RMSEP
PLS	10493	8	0.87	5.95	16.63	0.34	12.70
EN-PLS	309	7	0.93	4.34	6.93	0.55	11.66

R^2 is correlation coefficient of regression between the predicted and experimental activities of the extracts (t refers to training set and p refers to the test set); RMSET is the fitting error of the model in the training; RMSECV is the Root Mean Squared Errors of Cross-Validation; RMSEP is Root Mean Squared Errors of Prediction of the test set; q^2 is the cross-validated R^2 which is calculated by the equation: $q^2 = 1 - \sum(Y_{\text{pred}} - Y_{\text{act}})^2 / \sum(Y_{\text{act}} - Y_{\text{mean}})^2$.

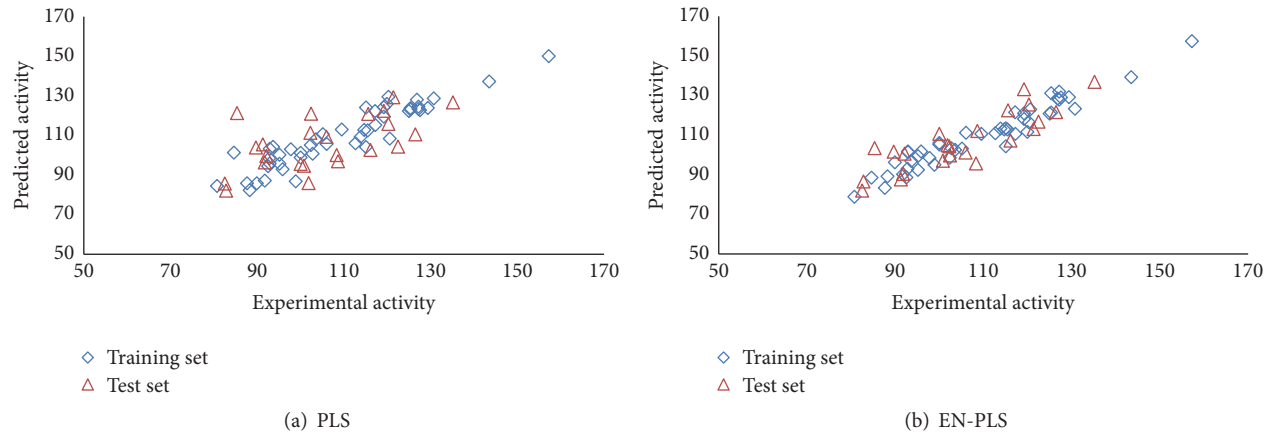


Figure 3: Plots of predicted versus experimental activity from training set data and test set data on (a) PLS and (b) EN-PLS. Open blue diamond and the open red triangle represent training set data and the test set data, respectively.

from batch A or batch B. We demonstrated that even the extracts of the same herb had different effects in modulating the CD80 expressions. This result indicated the diversity of the immunological effects as a result of the different chemical compositions of different extracts of the single herb RA.

3.6. Chemical and Biological Data Postprocessing and QPAR Model Development Using PLS. As discussed above, the chemical fingerprints of 72 RA extracts as an original dataset were described as data points or independent variables for construction of a model to get the relationship with their activities. Based on the Kennard and Stones algorithm [26], this original dataset was split into training set (48 samples, two-thirds of the total extracts) and external test set (24 samples, remaining one-third of the total extracts). Developed models were used to predict the CD80 immunomodulatory activity based on the chemical fingerprints provided in the test set. All the modeling analyses were carried out by MATLAB. To determine the degree of homogeneities of chemical fingerprints in the datasets, principle component analysis (PCA) was performed within the calculated descriptors space for all the chemical fingerprints.

Using the whole chromatographic retention time points as the variables, the first model was built by standard Partial Least Square (PLS). The PLS yielded a model having two correlation coefficients of regression values: Root Mean Squared Errors of Training (RMSRT) and Root Mean Squared Errors of Cross-Validation (RMSECV) of R_t^2 = 0.87 and RMSET = 5.95, respectively (Table 2). The PLS model had eight components with more than ten thousands variables.

3.7. Chemometric Model Refinement by EN-PLS. Due to the complexity of the chemical fingerprint with a large number of variables, further optimization by shrinkage methods was previously used to constrict the number of these variables. This optimization step was carried out to select those variables with high correlation with the biological activity. A selection method named Elastic Net (EN) was used to get a better predictive model [28]. To prove the ability of the model for QPAR study, internal cross-validation and external validation set (Test set) were applied to verify the predictability of the model (Supplementary Figure 4).

3.8. Quantitative Assessment of the QPAR Model Stability and Predictability. The QPAR models (training set) were built by PLS algorithms and the number of PLS components was determined by cross-validation. Figure 3 depicts the correlation regression figures of the experimental versus the predicted values for the training set data (open blue diamond) and the test set data (open red triangle) on PLS and EN-PLS models. From a leave-one-out cross-validation test applied to the training set, the best model, which gave the minimal sum value of the squared differences between predicted and experimental dependent variable, was determined.

The results obtained using PLS and EN-PLS for the training and the test sets were summarized in Table 3. The

Table 3: The actual CD80 activity (Act.) and the predicted values (Pred.) of the test set predicted by PLS and EN-PLS.

Number	Act	PLS			EN-PLS		
		Pred.	REP	PRESS	*Pred.*	REP	PRESS
1	108.33	100.04	−8.29	68.72	95.66	−12.67	160.53
2	105.95	109.26	3.31	10.96	101.16	−4.79	22.94
3	107.54	121.08	13.54	183.33	150.84	43.30	1874.89
4	100.00	95.54	−4.46	19.89	110.57	10.57	111.72
5	100.79	94.68	−6.11	37.33	97.03	−3.76	14.14
6	89.68	103.86	14.18	201.07	101.68	12.00	144.00
7	91.27	105.41	14.14	199.94	87.64	−3.63	13.18
8	85.32	121.30	35.98	1294.56	103.43	18.11	327.97
9	101.87	85.91	−15.96	254.72	104.86	2.99	8.94
10	102.25	111.26	9.01	81.18	104.34	2.09	4.37
11	108.61	96.89	−11.72	137.36	112.05	3.44	11.83
12	116.10	102.6	−13.50	182.25	107.22	−8.88	78.85
13	91.76	96.31	4.55	20.70	90.37	−1.39	1.93
14	92.13	99.56	7.43	55.20	100.61	8.48	71.91
15	82.83	82.06	−0.77	0.59	86.56	3.73	13.91
16	82.53	85.69	3.16	9.99	82.04	−0.49	0.24
17	115.58	120.84	5.26	27.67	122.50	6.92	47.89
18	120.29	115.73	−4.56	20.79	125.46	5.17	26.73
19	135.14	126.63	−8.51	72.42	136.89	1.75	3.06
20	126.45	110.55	−15.90	252.81	121.62	−4.83	23.33
21	122.46	104.37	−18.09	327.25	116.8	−5.66	32.04
22	121.38	129.29	7.91	62.57	113.11	−8.27	68.39
23	119.20	122.39	3.19	10.18	133.17	13.97	195.16
24	102.41	120.91	18.50	342.25	99.67	−2.74	7.51

REP = relative error of prediction = (calculated value-measured value)/measured value.
PRESS = predicted error sum of square for test set = $\sum(Y_{\text{pred}} - Y_{\text{act}})^2$.

performance of the model was firstly evaluated by R_t^2 that represented the correlation coefficient of regression between the fitted and experimental activities of the extracts in training set. In order to reflect the predictability power of a model, other parameters were used to avoid the overoptimistic error rate estimation and the model overfitting [29]. Both models demonstrated good fitting between predicted and experimental values in training set, where R_t^2 value was close to 0.9. The model built by EN-PLS ($R_t^2 = 0.93$) has better R_t^2 value compared with the standard PLS ($R_t^2 = 0.87$). Similarly, the model built by EN-PLS has better R_p^2 value of 0.55 when compared with the standard PLS ($R_p^2 = 0.34$, Table 2).

3.9. Computational Confirmation of the Predictability of the QPAR Models. Another quantitative measure of the stability and predictability of the PLS versus the EN-PLS was by comparing the RMSET and the RMSECV for the training set. The results showed that the EN-PLS model obtained the lowest values of RMSET and RMSECV of 4.34 and 6.93, respectively, in comparison to 5.95 and 16.63 for the standard PLS methodology (Table 2). For the test set of 24 samples, EN-PLS also generated a lower value of Root Mean Squared Errors of Prediction (RMSEP) of 11.66 in comparison with 12.70 when using the PLS (Table 2). In summary, PLS based on the Elastic Net variable selection method increased the biological predictive capability with lower value of RMSEP (11.66) and higher values of R_p^2 (0.55) when compared to the models developed by the standard PLS.

To provide further evidence that higher amount of predicted bioactive chemicals may induce corresponding biological CD80 expression, we first selected 13 regions from the chromatogram through detailed analyzed correlation coefficients of the PLS and EN-PLS models (Table 4). Six regions selected from the high positive correlation coefficients category, five regions from the high negative correlation coefficient category, and two regions from the zero correlation coefficient category have been selected. Based on the averaged chemical fingerprint of all 72 RA preparations, we increased the spectrophotometric intensities of each of these 13 selected regions (representing the amount of the specific compounds) by 50%, 100%, and 200% while keeping all other regions of the chromatogram unchanged, and the overall CD80 prediction was recalculated. Importantly, the results show that our model is able to predict correctly for both PLS and EN-PLS chemometrics approaches. All regions yielded a dose-response increase, decrease, or zero change in output according to their coefficient values. Furthermore, we

Table 4: The changes of CD80 expression related to % changes of chromatogram regions with different correlation coefficients.

Chromatogram region	Correlation coefficient generated by the prediction algorithm based on		% changes in CD80 expression when chromatogram region intensity (quantity of corresponding compound) was increased by 50, 100, or 200%					
			PLS			EN-PLS		
	PLS	EN-PLS	50% increase	100% increase	200% increase	50% increase	100% increase	200% increase
1	8.08	12.76	0.42	0.83	2.91	0.38	0.76	2.66
2	−10.59	−10.92	−0.98	−1.95	−3.90	−0.50	−1.01	−2.01
3	11.85	33.57	18.47	36.93	73.87	31.75	63.51	127.02
4	−9.07	−93.96	−1.19	−2.38	−4.76	−6.93	−13.86	−27.72
5	−10.29	−9.32	−0.14	−0.27	0.54	−0.06	−0.12	−0.25
6	9.07	30.64	0.14	0.28	0.56	0.23	0.47	0.93
7	−12.66	−55.19	−1.41	−2.82	−5.64	−3.30	−6.59	−13.18
8	−13.81	−56.59	−6.53	−13.05	−26.10	−14.00	−28.00	−56.00
9	10.63	41.34	0.62	1.25	2.50	1.34	2.68	5.36
10	12.19	48.64	0.15	0.31	0.61	0.34	0.68	1.37
11	−10.78	−30.46	−0.17	−0.35	−0.70	−0.24	−0.48	−0.95
12	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0

observed that, for one particular positive coefficient region, a 2-fold peak increase was related to a corresponding 127% CD80 expression increase. These calculations of changes of CD80 expression in relation to hypothetical modification of selected regions from the averaged chromatogram showed that higher amounts of bioactive chemicals induce stronger immune response.

4. Discussion

DCs play an important role in the regulation of tumor-specific immune responses [30]. However, cancer-associated microenvironment may adversely affect DC-related immune-surveillance system leading to defective DCs, which fail to upregulate important costimulatory surface molecule, CD80, and consequently ensue tumor escape and tolerogenicity [22, 23, 31]. According to the State Pharmacopoeia Commission of China, RA has been traditionally used in China to enhance human body's general well being. This effect on modulating the CD80 expression on THP-1 cells has been shown by our group [32].

In this study, we ultimately aimed to develop a predictive model of bioactivity for RA. Based on the CD80-QPAR approach, a model was built in association with the chemical compositions of the nonfractionated RA extract as represented by the fingerprint and the corresponding biological activity of CD80 expression modulation.

In our previous work, we used Target Projection (TP) to explore the bioactive components from a synthetic mixture system [33]. TP is good at eliminating "orthogonal variation" from inactive or weak bioactive components [34]. TP could reduce the QPAR model to a single component model based on an assumption if the total bioactivity is approximately additive in the bioactive molecular components.

However, if the total bioactivity of a whole extract is contributed also by interactions between molecules, that is, synergistic or antagonistic activities, it implies that the approach of reduction to a single predictive target component is no longer feasible. This study examining CD80 bioactivity of the RA extracts therefore adopted the PLS based on the Elastic Net variable selection method and considered that the overall sample bioactivities derived from the diverse chemical compositions of RA were contributed by each of the individual compounds as well as the multiple interactions between different compounds.

The EN model represents a useful grouping effect for model fitting and feature extraction, which selects those variables that have strong correlation with the bioactivity [28]. A regression model may exhibit the grouping effect when the regression coefficients of the highly correlated variables tend to be equal. In other words, the highly correlated variables will be selected. The performance and the predictability power of the EN-PLS were found to be superior to the conventional PLS methodology.

This study not only demonstrated the model's accurate predictability with the chromatogram alone of any new RA chemical preparations as input, it also facilitates greatly future drug discovery aiming to identify each of those components that contribute to these related CD80 expression modifications. In addition, future development of this CD80-QPAR platform should extend to the identification of those chemical compounds that presents in its native form to the metabolic derivatives [35]. Furthermore, this study sheds light on future laboratory studies on critical arenas of the synergistic [36] or antagonistic [37] effects in herbal mixtures and also the bioavailability and site-specificity issues [38, 39].

This study provides a clear illustration that AR may upregulate or downregulate the CD80 surface expression

of DC depending on different ways of preparations of RA, distinct compartments of the AR plant (batch B of outer part of RA versus batch C representing the inner core part of the RA), and different batches of RA purchased at different times. Our laboratory has previously shown that blood dendritic cells from patients with myeloma are numerically normal but functionally defective as they fail to upregulate CD80 (B7-1) expression after huCD40LT stimulation. This DC dysfunctionality is due to the high levels of inhibitory transforming growth factor-β1 and interleukin-10 in plasma [22, 23]. It is therefore important to understand that some RA preparations may have the desirable CD80 enhancement effect for cancer patients, whereas for autoimmunity patients RA preparations that have the biological effects of CD80 reduction are useful.

Other than the ability to affect the dendritic cells, the triterpene saponins extracted from RA have previously been shown to upregulate and activate T cells as shown by increased IL-2 production [40]. Some aqueous fractions of RA have shown to enhance allogenic T cell activity as shown by increased graft-versus-host reaction [41]. Furthermore, polysaccharides extracted from RA have shown to affect mouse B cells and macrophages but not the T cells [42]. Therefore, in future more bioactivity platforms of these key mechanisms of action of RA are required to have a more complete understanding of important compounds that are related to the overall immunomodulatory effects of RA.

5. Conclusions

In this CD80-QPAR study on a commonly used herb RA, we successfully explored and exploited the relationship between the chemical and biological fingerprints to establish a chemometric predictive model. Comparison between the statistical results, those obtained by Elastic Net variable selection method of Partial Least Square Method (EN-PLS), indicates the highest accuracy of QPAR study in describing the immunomodulatory activity of the ingredients from a commonly used food supplement of RA. PLS based on the Elastic Net variable selection method increased the biological predictive capability with lower value of RMSEP (11.66) and higher values of R_p^2 (0.55) when compared to the models developed by the standard PLS. The standard PLS approach can predict the CD80 bioactivity for unknown sample with an average of 10.05% difference; while the EN-PLS can predict the CD80 bioactivity with an average within only 7.59% difference, thus when using the EN selection method, there is a 25% improvement in the prediction capability.

With this CD80-QPAR platform, many herbal medicines in their entire crude extract without the need of tedious and time consuming immunomodulation bioactivity-guide fractionation can be screened for their bioactivities in moderating the CD80 expression using this robust THP-1 dendritic cell bioactivity platform. This study may bring novel insights into herbal vaccination-adjuvants preparation and may lead to correcting the defective dendritic cell CD80 costimulatory capacity. This paper also highlights the importance of how information technology may help the quality control process of the multiple components of the complex mixtures such as food supplements and herbal medicines for consistent batch-to-batch clinical usage in health and disease.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Michelle Chun-har Ng was responsible for data collection, data analysis, and manuscript preparation. Tsui-yan Lau performed data collection and data analysis. Kei Fan contributed conception and design. Josiah Poon, Simon K. Poon, and Mary K. Lam conducted critical discussion of the interpretation of chemometrics analytical methodology. Qing-song Xu performed data analysis and review of the manuscript. Foo-tim Chau contributed conception, design, data analysis, and review of the manuscript. Daniel Man-Yuen Sze contributed conception, design, data analysis, and review of the manuscript.

Acknowledgments

"The Fingerprint Analysis Software" was developed by the Research Centre of Modernization of Traditional Chinese Medicine of Central South University, Changsha, Hunan, China, in 2009. Daniel Man-Yuen Sze was supported by the National Institute of Complementary Medicine Collaborative Centre of Traditional Chinese Medicine Grant Australia. The authors thank also the expert bioinformatics technical support of Hao Chen and Alex Ng in generating Table 4 of this paper. The authors also acknowledge that Michelle Chun-har Ng has presented the preliminary data as an Abstract Presentation in the 15th International Congress of Immunology at Milan Italy during 22–27 August, 2013.

References

[1] K. Miller, B. Neilan, and D. M. Y. Sze, "Development of taxol and other endophyte produced anti-cancer agents," *Recent Patents on Anti-Cancer Drug Discovery*, vol. 3, no. 1, pp. 14–19, 2008.

[2] R. A. Khatib, M. Selemani, G. A. Mrisho et al., "Access to artemisinin-based anti-malarial treatment and its related factors in rural Tanzania," *Malaria Journal*, vol. 12, no. 1, article 155, 2013.

[3] L. Zhou, J. Hou, C. F. G. Chan, and D. M. Y. Sze, "Arsenic trioxide for non acute promyelocytic leukemia hematological malignancies: a new Frontier," *Journal of Blood Disorders*, vol. 1, no. 4, article 1018, 2014.

[4] W. Lam, S. Bussom, F. Guan et al., "The four-herb Chinese medicine PHY906 reduces chemotherapy-induced gastrointestinal toxicity," *Science Translational Medicine*, vol. 2, no. 45, Article ID 45ra59, 2010.

[5] S. Kummar, M. Sitki Copur, M. Rose et al., "A phase I study of the chinese herbal medicine PHY906 as a modulator of irinotecan-based chemotherapy in patients with advanced colorectal cancer," *Clinical Colorectal Cancer*, vol. 10, no. 2, pp. 85–96, 2011.

[6] T. S. Bugni, B. Richards, L. Bhoite, D. Cimbora, M. K. Harper, and C. M. Ireland, "Marine natural product libraries for high-throughput screening and rapid drug discovery," *Journal of Natural Products*, vol. 71, no. 6, pp. 1095–1098, 2008.

[7] S.-K. Yan, Z.-Y. Lin, W.-X. Dai et al., "Chemometrics-based approach to modeling quantitative composition-activity relationships for Radix Tinosporae," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 2, no. 3, pp. 221–227, 2010.

[8] C. Chen, S.-X. Li, S.-M. Wang, and S.-W. Liang, "A support vector machine based pharmacodynamic prediction model for searching active fraction and ingredients of herbal medicine: naodesheng prescription as an example," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 56, no. 2, pp. 443–447, 2011.

[9] J. L. Jiang, H. T. Ding, X. Su, and Y. J. Yuan, "Identification of anti-tumor ingredients in curcuma volatile oil based on composition-activity relationship," *Chinese Journal of Analytical Chemistry*, vol. 40, pp. 1488–1493, 2012.

[10] C. Tistaert, G. Chataigné, B. Dejaegher et al., "Multivariate data analysis to evaluate the fingerprint peaks responsible for the cytotoxic activity of Mallotus species," *Journal of Chromatography B*, vol. 910, pp. 103–113, 2012.

[11] J. Ching, W.-L. Soh, C.-H. Tan et al., "Identification of active compounds from medicinal plant extracts using gas chromatography-mass spectrometry and multivariate data analysis," *Journal of Separation Science*, vol. 35, no. 1, pp. 53–59, 2012.

[12] F.-T. Chau, H.-Y. Chan, C.-Y. Cheung, C.-J. Xu, Y. Liang, and O. M. Kvalheim, "Recipe for uncovering the bioactive components in herbal medicine," *Analytical Chemistry*, vol. 81, no. 17, pp. 7217–7225, 2009.

[13] F. T. Chau, Q. S. Xu, D. M. Y. Sze et al., "A new methodology for uncovering the bioactive fractions in herbal medicine using the approach of quantitative pattern-activity relationship," in *Data Analytics for Traditional Chinese Medicine Research*, pp. 155–172, Springer International, Berlin, Germany, 2014.

[14] D.-F. Wei, L.-F. Zhang, W.-D. Cheng et al., "[Comparative study of Hedysari Radix and Astragali Radix alternative classic tonification prescriptions on humoral immunity in immunosuppressed mice]," *Journal of Chinese Medicinal Materials*, vol. 35, no. 6, pp. 944–948, 2012.

[15] L.-F. Zhang, W.-D. Cheng, M.-M. Gui, X.-Y. Li, and D.-F. Wei, "Comparative study of Radix Hedyseri as sulstitute for *Radix Astragali* of yupingfeng oral liquid on cellular immunity in immunosuppressed mice," *Journal of Chinese Medicinal Materials*, vol. 35, no. 2, pp. 269–273, 2012.

[16] Y. Jung, U. Jerng, and S. Lee, "A systematic review of anticancer effects of Radix Astragali," *Chinese Journal of Integrative Medicine*, vol. 22, no. 3, pp. 225–236, 2016.

[17] Q. T. Gao, J. K. H. Cheung, J. Li et al., "A Chinese herbal decoction, Danggui Buxue Tang, activates extracellular signal-regulated kinase in cultured T-lymphocytes," *FEBS Letters*, vol. 581, no. 26, pp. 5087–5093, 2007.

[18] C.-C. Hsieh, W.-C. Lin, M.-R. Lee et al., "Dang-Gui-Bu-Xai-Tang modulated the immunity of tumor bearing mice," *Immunopharmacology and Immunotoxicology*, vol. 25, no. 2, pp. 259–271, 2003.

[19] J. Liu, X. Hu, Q. Yang et al., "Comparison of the immunoregulatory function of different constituents in radix astragali and radix hedysari," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 479426, 12 pages, 2010.

[20] J. Wang, X. Tong, P. Li, H. Cao, and W. Su, "Immunoenhancement effects of Shenqi Fuzheng Injection on cyclophosphamide-induced immunosuppression in Balb/c mice," *Journal of Ethnopharmacology*, vol. 139, no. 3, pp. 788–795, 2012.

[21] S. Kurashige, Y. Akuzawa, and F. Endo, "Effects of astragali radix extract on carcinogenesis, cytokine production, and cytotoxicity in mice treated with a carcinogen, N-butyl-N'- butanolnitrosoamine," *Cancer Investigation*, vol. 17, no. 1, pp. 30–35, 1999.

[22] R. D. Brown, B. Pope, A. Murray et al., "Dendritic cells from patients with myeloma are numerically normal but functionally defective as they fail to up-regulate CD80 (B7-1) expression after huCD40LT stimulation because of inhibition by transforming growth factor-β_1 and interleukin-10," *Blood*, vol. 98, no. 10, pp. 2992–2998, 2001.

[23] R. Brown, A. Murray, B. Pope et al., "Either interleukin-12 or interferon-γ can correct the dendritic cell defect induced by transforming growth factor β 1 in patients with myeloma," *British Journal of Haematology*, vol. 125, no. 6, pp. 743–748, 2004.

[24] W. K. Chan, H. K. W. Law, Z.-B. Lin, Y. L. Lau, and G. C.-F. Chan, "Response of human dendritic cells to different immunomodulatory polysaccharides derived from mushroom and barley," *International Immunology*, vol. 19, no. 7, pp. 891–899, 2007.

[25] W. K. Chan, C. C. H. Cheung, H. K. W. Law, Y. L. Lau, and G. C. F. Chan, "Ganoderma lucidum polysaccharides can induce human monocytic leukemia cells into dendritic cells with immuno-stimulatory function," *Journal of Hematology and Oncology*, vol. 1, no. 9, 2008.

[26] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[27] J.-Z. Song, S.-F. Mo, Y.-K. Yip, C.-F. Qiao, Q.-B. Han, and H.-X. Xu, "Development of microwave assisted extraction for the simultaneous determination of isoflavonoids and saponins in *Radix Astragali* by high performance liquid chromatography," *Journal of Separation Science*, vol. 30, no. 6, pp. 819–824, 2007.

[28] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[29] A. Golbraikh and A. Tropsha, "Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection," *Journal of Computer-Aided Molecular Design*, vol. 16, no. 5-6, pp. 357–369, 2002.

[30] R. M. Steinman, "The dendritic cell system and its role in immunogenicity," *Annual Review of Immunology*, vol. 9, no. 1, pp. 271–296, 1991.

[31] M. Y. Gerner, K. A. Casey, and M. F. Mescher, "Defective MHC class II presentation by dendritic cells limits CD4 T cell help for antitumor CD8 T cell responses," *The Journal of Immunology*, vol. 181, no. 1, pp. 155–164, 2008.

[32] M. C. H. Ng, K. Fan, T. Y. Lau, H. Y. Chan, F. T. Chau, and D. M. Y. Sze, "Investigation of the immunomodulatory effects of Radix Astragali targeting dendritic cells," in *Proceedings of the 11th International Symposium on Dendritic Cells in Fundamental and Clinical Immunology (DC '10): Forum on Vaccine Science*, Lugano , Switzerland, September 2010.

[33] O. M. Kvalheim, H.-Y. Chan, I. F. F. Benzie et al., "Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 98–105, 2011.

[34] J. Trygg and S. Wold, "Orthogonal projections to latent structures (O-PLS)," *Journal of Chemometrics*, vol. 16, no. 3, pp. 119–128, 2002.

[35] Y.-Z. Liang, P.-S. Xie, and K. Chan, "Chromatographic fingerprinting and metabolomics for quality control of TCM," *Combinatorial Chemistry and High Throughput Screening*, vol. 13, no. 10, pp. 943–953, 2010.

[36] H. Guo, H. Mao, G. Pan et al., "Antagonism of Cortex Periplocae extract-induced catecholamines secretion by Panax notoginseng saponins in cultured bovine adrenal medullary cells by drug combinations," *Journal of Ethnopharmacology*, vol. 147, no. 2, pp. 447–455, 2013.

[37] H.-Z. Yang, M.-M. Zhou, A.-H. Zhao, S.-N. Xing, Z.-Q. Fan, and W. Jia, "Study on effects of baicalin, berberine and Astragalus polysaccharides and their combinative effects on aldose reductase in vitro," *Journal of Chinese Medicinal Materials*, vol. 32, no. 8, pp. 1259–1261, 2009.

[38] J.-L. Huang, D.-P. Wu, L. Lu, F. Li, and Z.-G. Zhong, "The effect of PNS on the content and activity of alpha-secretase in the brains of SAMP8 mice with alzheimer's disease," *Journal of Chinese Medicinal Materials*, vol. 35, no. 11, pp. 1805–1808, 2012.

[39] T. Ma, K. Gong, Y. Yan et al., "Huperzine A promotes hippocampal neurogenesis in vitro and in vivo," *Brain Research*, vol. 1506, pp. 35–43, 2013.

[40] E. Yesilada, E. Bedir, İ. Çalış, Y. Takaishi, and Y. Ohmoto, "Effects of triterpene saponins from *Astragalus* species on in vitro cytokine release," *Journal of Ethnopharmacology*, vol. 96, no. 1-2, pp. 71–77, 2005.

[41] D.-T. Chu, W. L. Wong, and G. M. Mavligit, "Immunotherapy with Chinese medicinal herbs. I. Immune restoration of local xenogeneic graft-versus-host reaction in cancer patients by fractionated Astragalus membranaceus in vitro," *Journal of Clinical & Laboratory Immunology*, vol. 25, no. 3, pp. 119–123, 1988.

[42] B.-M. Shao, W. Xu, H. Dai, P. Tu, Z. Li, and X.-M. Gao, "A study on the immune receptors for polysaccharides from the roots of Astragalus membranaceus, a Chinese medicinal herb," *Biochemical and Biophysical Research Communications*, vol. 320, no. 4, pp. 1103–1111, 2004.

Hindawi
BioMed Research International
Volume 2017, Article ID 7653101, 7 pages
https://doi.org/10.1155/2017/7653101

Research Article

Prediction and Analysis of Key Genes in Glioblastoma Based on Bioinformatics

Hao Long,[1] Chaofeng Liang,[2] Xi'an Zhang,[1] Luxiong Fang,[1] Gang Wang,[1] Songtao Qi,[1] Haizhong Huo,[3] and Ye Song[1]

[1]*Department of Neurosurgery, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong 510515, China*
[2]*Department of Neurosurgery, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510665, China*
[3]*Department of General Surgery, Shanghai Ninth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200011, China*

Correspondence should be addressed to Haizhong Huo; huohz1570@sh9hospital.org and Ye Song; songye@smu.edu.cn

Received 1 September 2016; Accepted 21 November 2016; Published 16 January 2017

Academic Editor: Jens Schittenhelm

Understanding the mechanisms of glioblastoma at the molecular and structural level is not only interesting for basic science but also valuable for biotechnological application, such as the clinical treatment. In the present study, bioinformatics analysis was performed to reveal and identify the key genes of glioblastoma multiforme (GBM). The results obtained in the present study signified the importance of some genes, such as COL3A1, FN1, and MMP9, for glioblastoma. Based on the selected genes, a prediction model was built, which achieved 94.4% prediction accuracy. These findings might provide more insights into the genetic basis of glioblastoma.

1. Introduction

Glioblastomas are highly invasive tumors associated with high levels of mortality in the central nervous system, and their symptoms include bloating, pelvic pain, difficult eating, and frequent urination. It is difficult to diagnose glioblastoma at its early stages (I/II) as most symptoms of this disease are nonspecific [1]. Glioblastoma is a rare disease, with a rate of 2-3 cases per 100,000 person life-years in Europe and North America [2], accounting for 77–80% of primary malignant tumors of the brain. Among the patients diagnosed with glioblastoma, approximately 50% die within one year, while 90% die within three years [3]. Due to the great threat of glioblastoma to human health, the treatment of glioblastoma remains a major challenge.

Over the past years, tremendous genomics and proteomics studies have been conducted to explore the molecular mechanisms underlying the development and progression of glioblastoma. The characterization of glioblastoma has provided invaluable data related to this molecularly heterogeneous disease. Recent advances in high-throughput microarrays have received extensive attention and made substantial progress in reconstructing the gene regulatory network of medical biology [4–11]. Using microarray analysis, significant differences in gene expression between normal and disease tissues have been observed. However, as a result of the underlying shortcomings of microarray technology, such as small sample size, measurement error, and information insufficiency, unveiling this disease mechanism has remained a major challenge to glioblastoma research. Hence, GO, pathway information, network-based approaches, and machine learning algorithms have been employed to identify the mechanisms underlying this disease.

In the present study, we identified the differentially expressed genes (DEGs) between the glioblastoma samples and normal brain samples. In addition, eleven significant target genes for diagnosing glioblastoma were identified based on GO processes, KEGG pathways, and protein-protein interaction networks. Based on the results, a prediction model was built with a prediction accuracy of 94.4% with these eleven genes using Bayes net.

2. Materials and Methods

2.1. Data Preparation. The datasets available in this analysis contained 18 samples, including 9 glioblastoma tissue samples

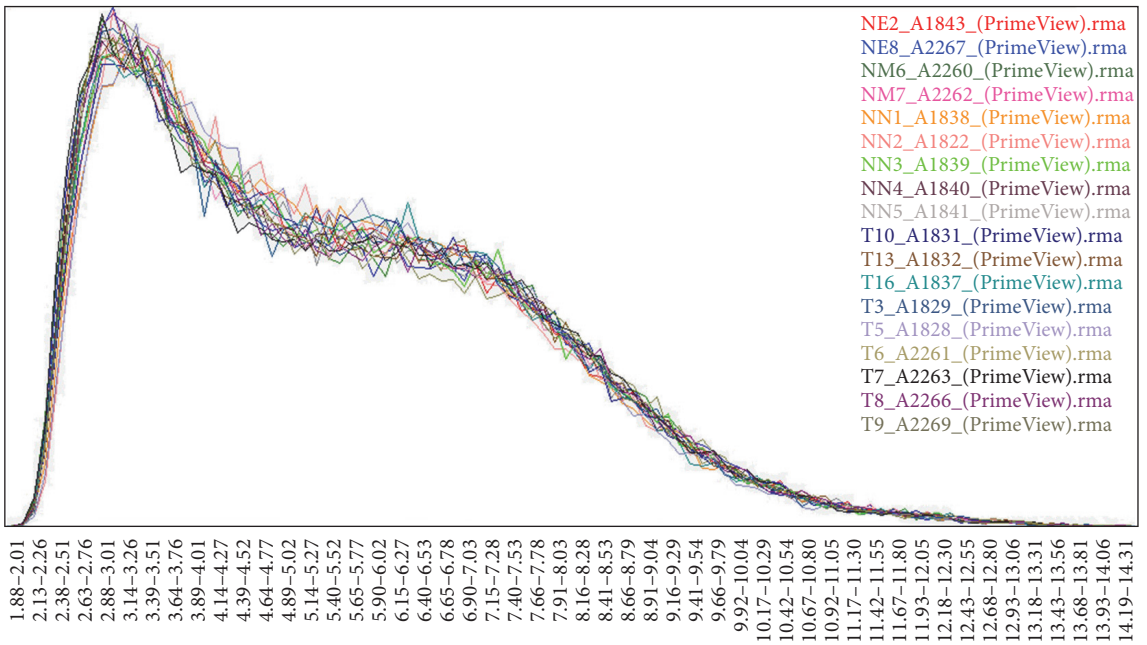


Figure 1: Histogram of the raw fluorescence intensity data.

and 9 normal brain tissue samples from epilepsy surgery. All specimens had confirmed pathological diagnosis and were classified according to the World Health Organization (WHO) criteria. All the tumor samples were obtained from primary surgery. For the use of these clinical materials for research purposes, prior consent from patients and approval from the Ethics Committees of Nanfang Hospital (number 2013105) were obtained. These data (CEL form) and annotation files were collected for further analysis. Figure 1 shows that the gene expression signals for the 18 samples fit well with each other and could be employed in the bioinformatics analysis in the present study.

3. Results

3.1. Raw Data. Limma package in R was used to identify the DEGs between the glioblastoma samples and the normal controls. According to the cut-off criteria of $|\log FC| > 2.0$ and p value < 0.05, we obtained 2365 DEGs, including 1021 up- and 1344 downregulated genes (please visit the following website for more raw data information: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90886).

3.2. Gene Ontology Analysis. GO analyses were performed by DAVID which demonstrated that the majority of DEGs were enriched in cellular components, cytoplasm, integral to membrane, intrinsic to membrane, biopolymer metabolic processes, cytoplasmic parts, and nucleus (Figure 2). The upregulated genes were significantly enriched in cytoplasm, nucleus, nucleobase-containing compounds, metabolic processes, and biopolymer metabolic processes.

3.3. Analysis of KEGG Pathways. To obtain further insight into the functions of DEGs, DAVID was applied to identify the significant dysregulated KEGG pathways. The pathways obtained with a p value < 0.05 and a gene count > 2 for the up- and downregulated genes were collected (Table 1). According to the enrichment results, the genes were significantly enriched in following pathways: cancer pathways, regulation of the actin cytoskeleton, the MAPK signaling pathway, focal adhesion, the calcium signaling pathway, ECM-receptor interaction, long-term potentiation, endocytosis, leukocyte transendothelial migration, and the p53 signaling pathway. Among these pathways, the upregulated genes were

Table 1: DEG pathway distribution.

KEGG pathway	DEGs	Upregulation	Downregulation
Calcium signaling pathway	45	4	41
MAPK signaling pathway	61	24	37
Endocytosis	37	11	26
Regulation of actin cytoskeleton	44	19	25
Long-term potentiation	24	2	22
Pathways in cancer	57	41	16
Focal adhesion	48	36	12
Leukocyte transendothelial migration	29	19	10
ECM-receptor interaction	33	29	4
p53 signaling pathway	22	22	0

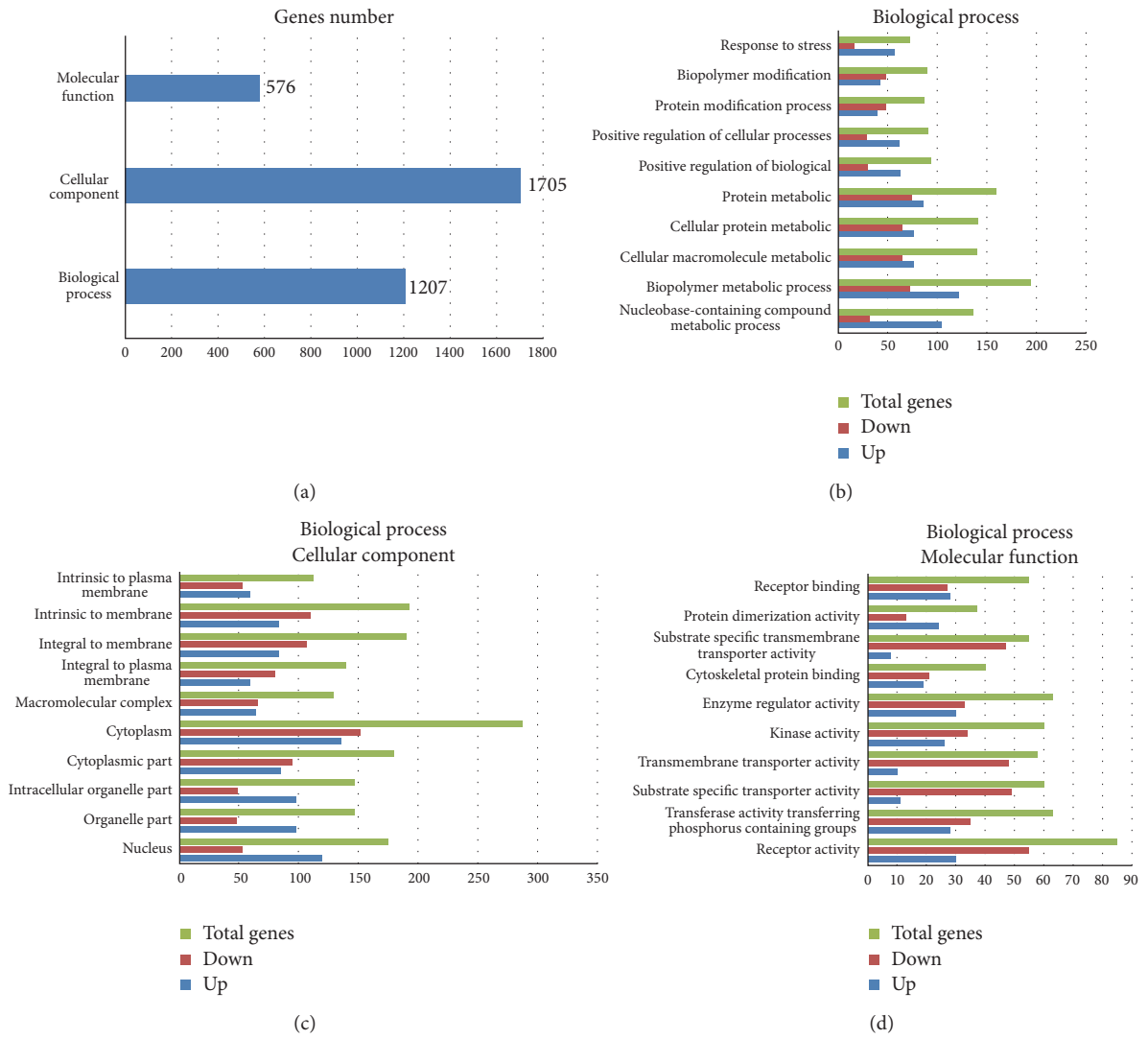


Figure 2: (a) GO enrichment of DEGs. (b) DEGs in BP. (c) DEGs in CC. (d) DEGs in MF.

significantly enriched in the pathways of focal adhesion, cancer, ECM-receptor interaction, MAPK signaling, and p53 signaling. The downregulated DEGs were enriched in the pathways of calcium signaling, MAPK signaling, endocytosis, regulation of actin cytoskeleton, and long-term potentiation.

3.4. PPI Network Construction. The STRING tool was used to determine the PPI relationships of the DEGs. In total, 2182 PPI relationships were obtained with a combined score >0.4. After filtering out the nodes of degree ≤5, we constructed a network with 240 nodes and 2182 edges (Figure 3(a)).

Based on the PPI network constructed above, PPI network enrichments were performed. The results revealed 5 enriched modules with a size >5 and a $p < 0.05$. Among the five modules, two significant enrichments, Module A and Module B, are shown in Figures 3(b) and 3(c). According to Figure 3(b), it is difficult to determine which module is better, as they had similar sizes and edges. However, as Module A has 38 nodes and 340 edges compared with Module B with 36 nodes and 320 edges, we considered Module A as the better module.

To investigate the biological functions of the genes in Module A, GO functional enrichments were performed using STRING tools. A total of 31 genes in Module A were significantly enriched in biological processes and cellular components, such as extracellular matrix organization, extracellular structure organization, extracellular region part, locomotion, and cell movement or subcellular components. Subsequently, these 31 genes were further investigated using KEGG pathway enrichment analysis. The results showed that the genes in Module A were primarily enriched by the following pathways: ECM-receptor interaction, focal adhesion, the PI3K-Akt signaling pathway, amoebiasis, protein digestion/absorption, and pathways in cancer.

The connectivity degree of each node of the PPI network was calculated, and the results of some nodes are shown in Table 2. As shown in Table 2, several genes, including MMP9, CD44, COL1A1, COL1A2, CAMK2A, and CAMK2B,

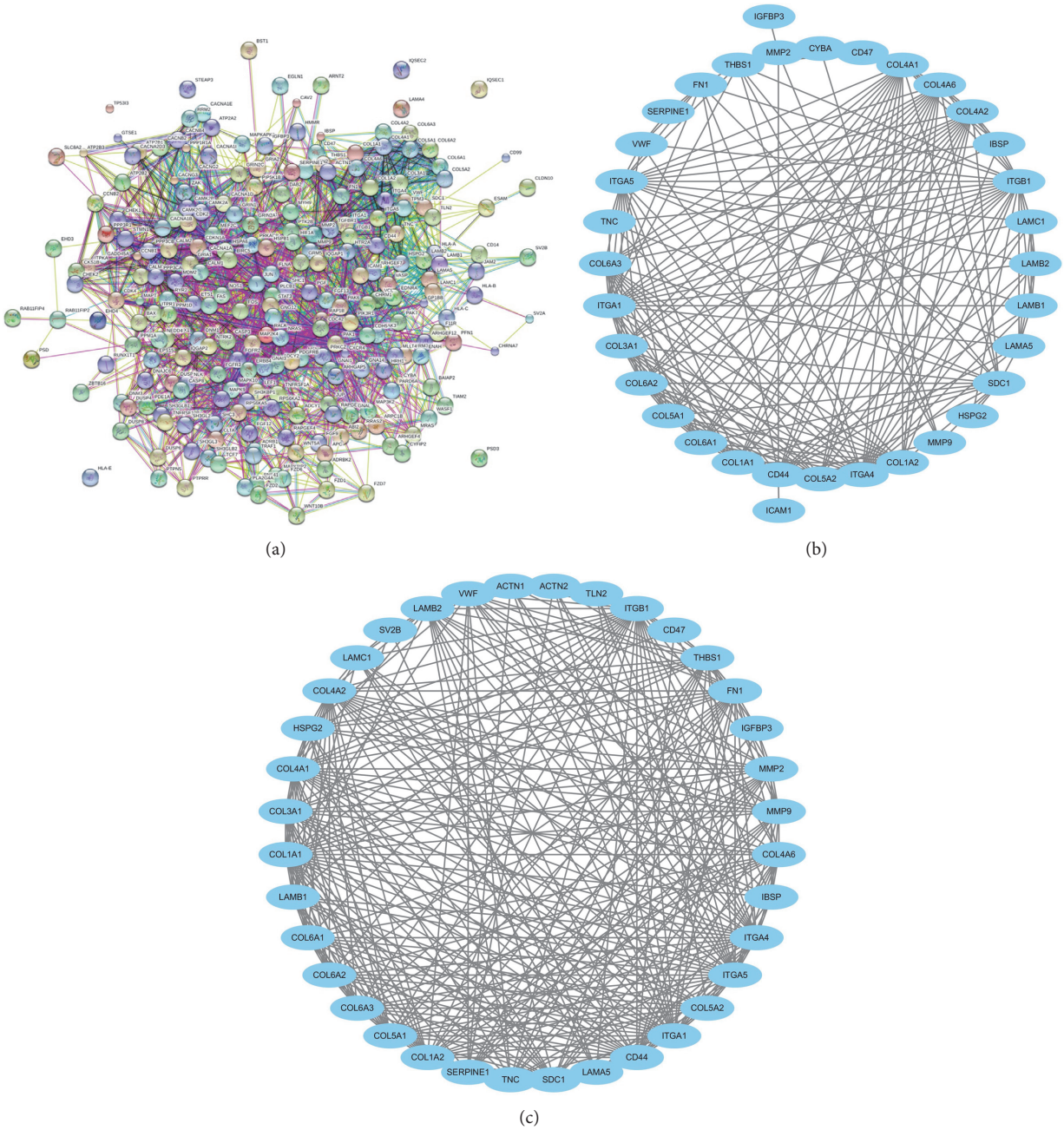


Figure 3: (a) Protein-protein interaction networks of the corresponding DEGs. ((b) and (c)) Modules of the PPI network.

exhibited a high connectivity degree >25. Hence, these genes were selected as key nodes and might play important roles in the progression of GBM.

3.5. Prediction Model. Based on the selected eleven genes, a predictive glioblastoma model was constructed using Bayes net algorithm. To validate the predictive capability of the model, a leave-one-out (LOO) cross-validation test, widely used in prediction-related problems, was adopted in the present study. For the LOO cross-validation test tests, the datasets were randomly divided into 18 subsets. Each classifier was constructed using the samples from seventeen of the subsets and the samples in the remaining subset were treated as untrained data, which were used in the prediction as independent test samples. Each subset was omitted when constructing the classifier and predicted in turn. The total prediction accuracy was obtained after averaging the correct prediction rates of the 18 data subsets. The following prediction results were obtained using the Bayes net method: SN: 88.9%, SP: 100%, ACC: 94.4%, and MCC: 0.795.

4. Discussion

In the present study, we obtained 2365 genes, including 1021 upregulated genes and 1344 downregulated genes using gene expression profiling. Among the 2365 genes identified,

Table 2: The statistical results of the connectivity degrees of the PPI network.

Gene	Degree	Differential rates
CDC42	73	−6.928589
MMP9	49	16.665218
CD44	41	14.821273
CAV1	39	5.418803
THBS1	35	6.9339356
CAMK2B	33	−7.2954154
CAMK2A	32	−14.86015
COL1A2	30	9.719963
FN1	28	9.536128
COL4A2	27	7.803446
COL3A1	26	27.03572

there were 365 differentially expressed genes, including 237 upregulated genes and 124 downregulated genes. Most of these genes were enriched in ten pathways, including MAPK signaling, cancer, focal adhesion, calcium signaling, actin cytoskeleton regulation, endocytosis, ECM-receptor interaction, leukocyte transendothelial migration, long-term potentiation, and p53 signaling pathways. Moreover, the upregulated DEGs were primarily enriched in pathways in cancer, focal adhesion, and ECM-receptor interaction, while the downregulated DEGs were significantly related to pathways, such as the calcium signaling pathway, MAPK signaling pathway, and endocytosis. COL3A1, MMP9, CAMK2A, CD44, HTR2A, SV2B, GRIN2A, COL6A3, and SH3GL3 have been identified as significant genes in these pathways. MMP9, FN1, FGF13, and COL4A2 are significant genes in the pathways associated with cancer. COL3A1, COL6A3, COL1A2, FN1, and TNC are significant genes in the focal adhesion pathway. CAMK2A, HTR2A, and GRIN2A are significant genes in the calcium signaling pathway. COL3A1, CD44, SV2B, and COL6A3 are significant genes in ECM-receptor interactions.

These results indicate that the ECM-receptor interaction pathway is a significant pathway enriched by upregulated DEGs. In the present study, COL3A1 and CD44 in ECM-receptor interaction pathway were significantly upregulated. CD44, an unclassified cell adhesion molecule, is involved in cell-cell interactions, cell adhesion, and migration [12, 13]. Studies have shown that CD44 participates in a wide variety of cellular functions, including lymphocyte activation and the recirculation, recurrence, and development of tumors [14]. In a previous study, Yoshida indicated that the overexpression of CD44 was important for the growth and survival of glioblastomas, and the monoclonal anti-CD44 antibody affects the migration of glioblastoma cells [15, 16]. COL3A1 encodes fibrillar collagen, a major component of the extracellular matrix protein surrounding cancer cells [17, 18]. The presence of ECM protein prevents the apoptosis of cancer cells. COL3A1 plays an important role in apoptosis, proliferation regulation, and anticancer drug resistance [19], indicating that the ECM-receptor interaction pathway plays an important role in GBM, and CD44 and COL3A1 might be potential diagnostic and therapeutic targets in this disease.

In the present study, MMP9 and FN1, key proteins in cancer pathways, were also upregulated. The proteins of the matrix metalloproteinase (MMP) family are involved in the breakdown of extracellular matrix in normal biological processes, such as embryonic development, angiogenesis, cell migration, intracerebral hemorrhage, and metastasis [20, 21]. As a member of the MMPs, MMP9 is involved in the degradation of the extracellular matrix. MMP9 also plays roles in tumor development, as these proteins facilitate extracellular matrix remodeling and participate in angiogenesis. Forsyth et al. reported the involvement of MMP9 in different aspects of the pathophysiology of malignant gliomas by remodeling associated with neovascularization [22]. Choe et al. detected MMP9 in the tumor samples of GBM patients but not in normal brain tissue samples. Moreover, these authors also showed that EGFRvIII overexpression affects MMP9 activation by the activation of MAPK/ERK [23]. FN1, a high-molecular weight glycoprotein of the extracellular matrix, binds extracellular matrix components, such as collagen, fibrin, and heparan sulfate proteoglycans. Wang et al. reported that FN is involved in the maintenance of integrin b1 fibronectin receptors in glioma cells and could be regarded as an important mediator [24]. Han et al. proposed that fibronectin stimulates non-small cell lung carcinoma cell growth and survival through the activation of the Akt/mTOR/p70S6K pathway [25], and recently, fibronectin has been implicated in carcinoma development as a potential biomarker for radioresistance [14].

Yu and Stamenkovic identified a functional relationship between the hyaluronan receptor CD44, MMP9, and transforming growth factor-beta in the control of tumor-associated tissue remodeling [26, 27]. These authors also showed that several isoforms of CD44, expressed on murine mammary carcinoma cells, provide cell surface docking receptors for proteolytically active MMP9. The localization of MMP9 on the cell surface is required to promote tumor invasion and angiogenesis. Moreover, the cell surface expression of MMP9 stimulated the formation of capillary tubes by bovine microvascular endothelial cells.

5. Conclusions

The results of the present study suggested that glioblastoma is closely associated with the dysregulation of the pathways in cancer, MAPK signaling, focal adhesion, and calcium signaling. In addition, we also identified key genes, including MMP9, CD44, CDC42, COL1A1, COL1A2, CAMK2A, and CAMK2B, as potential target genes for diagnosing glioblastoma.

Disclosure

The funders had no role in study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contributions

Hao Long and Chaofeng Liang have contributed equally to this work.

Acknowledgments

This study was supported by National Natural Science Foundation of China (nos. 81372692, 81502178, and 81502177) (http://www.nsfc.gov.cn), Fund of Development Center for Medical Science and Technology National Health and Family Planning Commission of China (no. W2013FZ15) (http://www.dcmst.org.cn/), Natural Science Foundation of Guangdong Province (nos. 2014A030313303, 2014A030313282, and 2016A030313549) (http://www.gdstc.gov.cn), Science and Technology Project of Guangdong Province (no. 2013B021800086) (http://www.gdstc.gov.cn), and President Fund of Nanfang Hospital (2013Z008 and 2014B007) (http://www.nfyy.com).

References

[1] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer Genetics*, vol. 205, no. 12, pp. 613–621, 2012.

[2] F. E. Bleeker, R. J. Molenaar, and S. Leenstra, "Recent advances in the molecular understanding of glioblastoma," *Journal of Neuro-Oncology*, vol. 108, no. 1, pp. 11–27, 2012.

[3] CBTRUS in CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2004–2006. Central Brain Tumor Registry of the United States, Hinsdale, Ill, USA, 2010, http://www.cbtrus.org/.

[4] J. Kononen, L. Bubendorf, A. Kallioniemi et al., "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine*, vol. 4, no. 7, pp. 844–847, 1998.

[5] C. Bucher, J. Torhorst, L. Bubendorf et al., "Tissue microarrays ('tissue chips') for high-throughput cancer genetics: linking molecular changes to clinical endpoints," *American Journal of Human Genetics*, vol. 65, no. 4, p. A10, 1999.

[6] R. Radhakrishnan, M. Solomon, K. Satyamoorthy, L. E. Martin, and M. W. Lingen, "Tissue microarray—a high-throughput molecular analysis in head and neck cancer," *Journal of Oral Pathology & Medicine*, vol. 37, no. 3, pp. 166–176, 2008.

[7] C. M. Kelly, S. Penny, D. Brennan et al., "Systematic validation of novel breast cancer progression-associated biomarkers *via* high-throughput antibody generation and application of tissue microarray technology: an initial report," *Journal of Clinical Oncology*, vol. 26, no. 15, supplement, p. 11056, 2008.

[8] T. G. Fernandes, S. J. Kwon, M. Y. Lee, D. S. Clark, J. M. S. Cabral, and J. S. Dordick, "On-chip, cell-based microarray immunofluorescence assay for high-throughput analysis of target proteins," *Analytical Chemistry*, vol. 80, no. 17, pp. 6633–6639, 2008.

[9] M. Izumiya, K. Okamoto, N. Tsuchiya, and H. Nakagama, "Functional screening using a microRNA virus library and microarrays: a new high-throughput assay to identify tumor-suppressive microRNAs," *Carcinogenesis*, vol. 31, no. 8, pp. 1354–1359, 2010.

[10] J.-H. Rho and P. D. Lampe, "High-throughput screening for native autoantigen-autoantibody complexes using antibody microarrays," *Journal of Proteome Research*, vol. 12, no. 5, pp. 2311–2320, 2013.

[11] M. G. Dozmorov and J. D. Wren, "High-throughput processing and normalization of one-color microarrays for transcriptional meta-analyses," *BMC Bioinformatics*, vol. 12, supplement 10, article S2, 2011.

[12] T. E. I. Taher, R. van der Voort, L. Smit et al., "Cross-talk between CD44 and c-met in B cells," in *Mechanisms of B Cell Neoplasia 1998: Proceedings of the Workshop held at the Basel Institute for Immunology 4th–6th October 1998*, F. Melchers and M. Potter, Eds., vol. 246 of *Current Topics in Microbiology and Immunology*, pp. 31–38, Springer, Berlin, Germany, 1999.

[13] G. F. Weber, S. Ashkar, M. J. Glimcher, and H. Cantor, "Receptor-ligand interaction between CD44 and osteopontin (Eta-1)," *Science*, vol. 271, no. 5248, pp. 509–512, 1996.

[14] D. Naor, R. V. Sionov, and D. Ish-Shalom, "CD44: structure, function, and association with the malignant process," *Advances in Cancer Research*, vol. 71, pp. 241–319, 1997.

[15] H. Okada, J. Yoshida, M. Sokabe, T. Wakabayashi, and M. Hagiwara, "Suppression of CD44 expression decreases migration and invasion of human glioma cells," *International Journal of Cancer*, vol. 66, no. 2, pp. 255–260, 1996.

[16] T. Yoshida, Y. Matsuda, Z. Naito, and T. Ishiwata, "CD44 in human glioma correlates with histopathological grade and cell migration," *Pathology International*, vol. 62, no. 7, pp. 463–470, 2012.

[17] U. Schwarze, W. I. Schievink, E. Petty et al., "Haploinsufficiency for one COL3A1 allele of type III procollagen results in a phenotype similar to the vascular form of Ehlers-Danlos syndrome, Ehlers-Danlos syndrome type IV," *American Journal of Human Genetics*, vol. 69, no. 5, pp. 989–1001, 2001.

[18] L. S. Payne and P. H. Huang, "The pathobiology of collagens in glioma," *Molecular Cancer Research*, vol. 11, no. 10, pp. 1129–1140, 2013.

[19] J. Skog, T. Würdinger, S. van Rijn et al., "Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers," *Nature Cell Biology*, vol. 10, no. 12, pp. 1470–1476, 2008.

[20] J. Wang and S. E. Tsirka, "Neuroprotection by inhibition of matrix metalloproteinases in a mouse model of intracerebral haemorrhage," *Brain*, vol. 128, pp. 1622–1633, 2005.

[21] J. Vandooren, P. E. Van den Steen, and G. Opdenakker, "Biochemistry and molecular biology of gelatinase B or matrix metalloproteinase-9 (MMP-9): the next decade," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 48, no. 3, pp. 222–272, 2013.

[22] P. A. Forsyth, H. Wong, T. D. Laing et al., "Gelatinase-A (MMP-2), gelatinase-B (MMP-9) and membrane type matrix metalloproteinase-1 (MT1-MMP) are involved in different aspects of the pathophysiology of malignant gliomas," *British Journal of Cancer*, vol. 79, no. 11-12, pp. 1828–1835, 1999.

[23] G. Y. Choe, J. K. Park, L. Jouben-Steele et al., "Active matrix metalloproteinase 9 expression is associated with primary glioblastoma subtype," *Clinical Cancer Research*, vol. 8, no. 9, pp. 2894–2901, 2002.

[24] F. F. Wang, G. Song, M. Liu, X. Li, and H. Tang, "miRNA-1 targets fibronectin1 and suppresses the migration and invasion of the HEp2 laryngeal squamous carcinoma cell line," *FEBS Letters*, vol. 585, no. 20, pp. 3263–3269, 2011.

[25] S. W. Han, F. R. Khuri, and J. Roman, "Fibronectin stimulates non-small cell lung carcinoma cell growth through activation of Akt/mammalian target of rapamycin/S6 kinase and inactivation of LKB1/AMP-activated protein kinase signal pathways," *Cancer Research*, vol. 66, no. 1, pp. 315–323, 2006.

[26] Q. Yu and I. Stamenkovic, "Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-beta and promotes tumor invasion and angiogenesis," *Genes & Development*, vol. 14, no. 2, pp. 163–176, 2000.

[27] Q. Yu and I. Stamenkovic, "Transforming growth factor-beta facilitates breast carcinoma metastasis by promoting tumor cell survival," *Clinical & Experimental Metastasis*, vol. 21, no. 3, pp. 235–242, 2004.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4596326, 10 pages
http://dx.doi.org/10.1155/2016/4596326

Research Article

Random Subspace Aggregation for Cancer Prediction with Gene Expression Profiles

Liying Yang,[1] Zhimin Liu,[1] Xiguo Yuan,[1] Jianhua Wei,[2] and Junying Zhang[1]

[1]*School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China*
[2]*State Key Laboratory of Military Stomatology, Department of Maxillofacial Surgery, School of Stomatology, the Fourth Military Medical University, Xi'an, China*

Correspondence should be addressed to Liying Yang; yangliying1208@163.com and Jianhua Wei; weiyoyo@fmmu.edu.cn

Received 3 July 2016; Revised 8 October 2016; Accepted 20 October 2016

Academic Editor: Bing Niu

Background. Precisely predicting cancer is crucial for cancer treatment. Gene expression profiles make it possible to analyze patterns between genes and cancers on the genome-wide scale. Gene expression data analysis, however, is confronted with enormous challenges for its characteristics, such as high dimensionality, small sample size, and low Signal-to-Noise Ratio. *Results*. This paper proposes a method, termed RS_SVM, to predict gene expression profiles via aggregating SVM trained on random subspaces. After choosing gene features through statistical analysis, RS_SVM randomly selects feature subsets to yield random subspaces and training SVM classifiers accordingly and then aggregates SVM classifiers to capture the advantage of ensemble learning. Experiments on eight real gene expression datasets are performed to validate the RS_SVM method. Experimental results show that RS_SVM achieved better classification accuracy and generalization performance in contrast with single SVM, K-nearest neighbor, decision tree, Bagging, AdaBoost, and the state-of-the-art methods. Experiments also explored the effect of subspace size on prediction performance. *Conclusions*. The proposed RS_SVM method yielded superior performance in analyzing gene expression profiles, which demonstrates that RS_SVM provides a good channel for such biological data.

1. Introduction

Cancer usually has an association with genes which carry human heritage information. Completion of human genome sequencing makes genetic analysis on the genome-wide scale available and provides a deeper understanding of the underlying mechanism of cancers [1–4]. Biological technology now can simultaneously monitor ten thousands of gene expression levels [5, 6]. It is meaningful to design novel methods to precisely and efficiently classify tumor samples from normal samples or recognize subclasses of some disease with gene expression profiles. Classification of gene expression data, however, faces enormous difficulties. Firstly, the data have up to ten thousands of dimensions. Traditional classification methods become intractable, since high dimensionality makes sample distribution dispersing and distance between samples ambiguous. Secondly, sample size is small for high expenses or ethical consideration. Therefore, there is no enough data to train a classical learner. Low Signal-to-Noise Ratio (SNR) is the third issue to consider for gene expression data analysis, which means noise may significantly decline performance.

To tackle the high dimensionality issue, some researches make an attempt to select important gene features by exploiting the association among genes and eliminating redundant and irrelevant information. Based on Recursive Feature Elimination (RFE), Guyon et al. used SVM method to select genes and proved that the genes filtered by SVM method perform better [7]. By feature extraction and defining "correlation feature space" for samples built on gene expression profiles through iterative utilization of Pearson's correlation coefficient, Ren et al. proposed an original method to further propel gene expression profiling technologies from bench to bedside [8]. Considering the possible interactions among

genes, Zhang et al. proposed a binary matrix shuffling filter to surmount troubles linked with searching schemes of conventional wrapper method and overfitting [9].

Ensemble art is also introduced in some recent researches. Bolón-Canedo et al. provided a novel framework for feature selection by an ensemble of filters and classifiers [10]. Combining classifiers from different classification families into an ensemble based on the evaluation of performance of each classifier, Nagi and Bhattacharyya proposed an ensemble method named as SD-EnClass [11]. To ensure a high classification accuracy, Ghorai et al. showed an ensemble of nonparallel plane proximal classifiers based on the genetic algorithm through simultaneous feature and model selection scheme [12]. Given the fact that forward feature selection (FFS) method is able to obtain an expected feature subset with less iteration than that of backward feature selection (BFS) method, Luo et al. proposed two FFS methods based on the pruning of the classifier ensembles generated by a single gene feature [13].

"Blessing of nonuniformity" effect, which means samples are concentrated in a relatively low instance space rather than uniformly throughout the whole space, inspired some novel methods to perform classification in subspaces [14]. Constructing subspace in random process was firstly proposed by Ho for decision forests to overcome the dilemma between avoiding overfitting and achieving maximum accuracy [15].

Recently, researchers have done much work on cancer classification based on gene expression data. Daxa et al. proposed a framework to find informative gene combinations and to classify gene combinations belonging to their relevant subtype by using fuzzy logic, while they only focused on identifying 2-gene and 3-gene combinations [16]. Kim et al. presented a genetic filter to identify gene subset for cancer-type classification on gene expression profiles, which was only tested on one dataset, that is, Leukemia dataset [17]. Vosooghifard and Ebrahimpour proposed a hybrid method using GWO and C4.5 for gene selection and cancer classification. In essence, GWO is a group optimization method, so time consuming is a factor which should be considered [18]. Buza summarized the classification of gene expression data in reference [19], where he indicated that the robustness of SVM to classify gene expression data relies on the strong fundamentals of statistical learning theory.

This paper attempts to classify gene expression data by aggregating SVMs trained on random subspaces (RS). RS method shows great potential in scenarios where the number of features is much bigger than the number of samples [20–23]. In addition, RS method has an excellent performance in coping with correlation and redundancy between features. Bias risk is relatively small in RS because of its independence of specific hypothesis on data. SVM is usually used to cope with gene expression data, since only support vectors work in classification process, and the number of support vectors is usually much smaller than that of training samples. We elaborately explored the trick of choosing parameters and the effect of size of subspaces on the classification performance. The possible reason leading to unsatisfied outcome was also revealed.

2. Materials and Methods

2.1. Gene Expression Datasets. Eight real gene expression datasets are used. They are provided by Kent Ridge Biomedical Dataset Repository and collected by Li and Liu from Nanyang Technological University, Singapore [24]. Detailed information is listed in Table 1.

Breast Cancer dataset labels the patients who had got distance metastases in five years as "relapse" and label the patients who remained healthy since the initial diagnosis for interval of at least five years as "nonrelapse." Missing values are replaced by 100 [25].

Leukemia dataset was originally published in reference [26]. Dataset used in this work is an extended and more heterogeneous version than the initial one. Samples are from DFCI (Dana-Farber Cancer Institute), CALGB (Cancer and Leukemia Group B), and SJCRH (St. Jude Children's Research Hospital). There are two categories, ALL (Acute Lymphoblastic Leukemia) and AML (Acute Myeloid Leukemia), inside the total 72 samples over 7129 probes. Training dataset consists of 38 bone marrow samples (27 ALL and 11 AML), while 34 testing samples (20 ALL versus 14 AML) are provided with 24 bone marrow and 10 peripheral blood specimens.

Lung Cancer dataset was firstly presented in reference [27]. Training set consists of 16 malignant pleural mesothelioma (MPM) samples and 16 adenocarcinoma (ADCA) samples. Testing set contains 15 MPM samples and 134 ADCA samples. 12533 genes expression levels were obtained via hybridizing cRNA to human U95A oligonucleotide probe arrays. All the ADCA samples and 12 MPM samples were processed at the Dana-Farber Cancer Institute and the Whitehead Institute. The remaining 19 MPM samples were processed separately at Brigham and Women's Hospital.

Prostate dataset has an independent testing set, which is from a different experiment and has a nearly tenfold difference in overall microarray intensity from the training data [40].

Colon Tumor dataset was introduced in reference [41]. Rather than elaborating time-course data, this dataset consists of snapshots of the expression pattern of distinct cell types. Raw dataset, based on 22 normal colon tissue samples (positive) and 40 colon tumor samples (negative) of colon adenocarcinoma specimens, was from an Affymetrix oligonucleotide array complementary to more than 6,500 genes and expressed sequence tags (ESTs). Two thousand genes were selected to generate the dataset used here, with the highest minimal intensity across 62 samples.

CNS (central nervous system) dataset was originally published in reference [42], while only dataset C mentioned to analyze the outcome of the treatment is used here. 60 samples consist of 39 medulloblastoma survivors (Class 0) and 21 treatment failures (Class 1). The dataset contains 60 patient samples, with 21 medulloblastoma survivors (labelled as "Class 1") and 39 treatment failures (labelled as "Class 0"). There are 7129 genes in the dataset.

Ovarian dataset was originally published in reference [43], inside which experiments are to identify proteomic patterns in serum that distinguish ovarian cancer from non-

Table 1: Dataset.

Data	Feature	Sample	Class
Breast Cancer	24481	97 78 training (34 relapse + 44 nonrelapse) 19 test (12 relapse + 7 nonrelapse)	Relapse Nonrelapse
Leukemia	7129	72 38 training (27 ALL + 11 AML) 34 test (20 ALL + 14 AML)	All AML
Lung Cancer	12533	181 32 training (16 mesothelioma + 16 ADCA) 149 test (15 mesothelioma + 134 ADCA)	Mesothelioma ADCA
Prostate	12600	136 102 training (52 tumor + 50 normal) 34 test (25 tumor + 9 normal)	Tumor Normal
Colon Tumor	2000	62 22 positive + 40 negative	Positive Negative
CNS	7129	60 21 Class 1 + 39 Class 0	Class 1 Class 0
Ovarian	15154	253 162 cancer + 91 normal	Cancer Normal
DLBCL	4026	47 24 germinal + 23 activated	Germinal Activated

cancer. The proteomic spectra were generated by mass spectroscopy and dataset used in this work includes 91 "Normal" samples and 162 "Cancer" samples without separated training set and testing set. The raw spectral data of each sample contains the relative amplitude of the intensity at each molecular mass/charge (M/Z) identity. There are totally 15154 M/Z identities. The intensity values were normalized according to the formula $\mathrm{NV} = (V - \mathrm{Min})/(\mathrm{Max} - \mathrm{Min})$, where NV is the normalized value, V the raw value, Min the minimum intensity, and Max the maximum intensity. The normalization is done over all the 253 samples for all 15154 M/Z identities. Thus, each intensity value falls into the range of 0 to 1.

As the most common subtype of non-Hodgkin's lymphoma, DLBCL (diffuse large B cell lymphoma) is due to an aggressive malignancy of mature B lymphocytes. DLBCL consists of two molecularly different subclasses [44]. One subclass is "germinal centre B like DLBCL" expressing gene characteristics of germinal centre B cells and the other is "activated B-like DLBCL" expressing genes normally induced during *in vitro* activation of peripheral blood B cells. DLBCL dataset contains 47 mRNA samples consisting of 24 germinal centre B-like DLBCL and 23 activated B-like DLBCL. Each of 4026 column score responding to cDNA clones indicates a gene expression level. Log-transformation was implemented on raw dataset to produce the dataset used in this work.

2.2. Method Description. SVM has an advantage in small sample cases and RS method shows an excellent performance in coping with high-dimension data. Algorithm 1 presents a description of RS_SVM method used in this paper, which aggregates SVMs trained on random subspaces. Figure 1 shows the framework of RS_SVM.

2.3. Gene Selection. Gene expression profile usually contains a large number of genes with constant or near constant expression levels across samples. These genes are redundant for classification and even decline distinction between normal and tumor samples, since they sharply increase space dimensions. To address this problem, gene selection based on statistical analysis is adopted to yield a new gene set from the original one. Since t-test is the first method for feature selection when microarray technology came into being, it is used in this work. Firstly, we compute p value of each gene across total samples and rank genes according to p value; then, top genes are filtered at 0.95 significant level. Number of top genes and optimal size of subspace on eight datasets are presented in Table 2.

2.4. Size and Number of Random Subspaces. Random subspace size (S) has an enormous influence on RS_SVM. Supposing that S value is relatively small, some important gene features may not be selected into feature subsets to train SVMs; thus, underfitting easily occurs. In contrast, if S is extremely large, diversity among SVM classifiers may be reduced, leading to a useless aggregation. Following experiment sets, default S to be the square root of M (feature number of selected data by t-test), recommended by Breiman [45], and then adjust S until achieving the optimal testing error. We analyze the influence of random subspace size on classification performance via illustrating the variation of training error and testing error with different S in Figure 3. An appropriate number of random subspaces (L) can guarantee that each feature has enough chance to be selected. Since the lack of prior knowledge about L, it is set to 1000 experimentally.

Input:
Dataset $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$, sample size n;
Sample $x^{(i)} = \{x_1{}^{(i)}, x_2{}^{(i)}, \ldots, x_m{}^{(i)}\}$, number of total feature m;
Class of ith sample $y^{(i)}$ in $Y = \{\text{normal, tumor}\}$;
Split function: yield training set and testing set from original dataset. If the original dataset has been divided into training and testing partition, this step could be skipped.
Gene select function: $\mathbf{R}^m \rightarrow \mathbf{R}^M$, where M is the feature number of selected data, $M < m$;
RS_preject function: $\mathbf{R}^M \rightarrow \mathbf{R}^S$, where S is the size of a random subspace, $S < M$;
Number of random subspaces L;
Learning algorithm: SVM
Output:
Classification hypotheses H: $X \rightarrow Y$
Start:
Data processing:
(Trainset, Testset) = Split(D)
TrainsetNew = Gene_select(Trainset, M)
TestsetNew = Gene_select(Testset, M)
Generate and aggregate SVM classifiers on random subspaces:
For $i = 1$ to L
D_i = RS_project(TrainsetNew, S)
h_i = SVM(D_i)
End
Test:
For each x in TestsetNew
$H(x) = \arg\max_{y \in Y} \sum_{i=1}^{L} (h_i(x) = y)$
End
End

Algorithm 1

Table 2: Number of selected features and optimal size of subspace.

Data	Number of selected features by t-test	Optimal size of subspace
Breast Cancer	1810	800
Leukemia	1697	400
Lung Cancer	3134	170
Prostate	5707	100
Colon Tumor	394	150
CNS	378	180
Ovarian	7949	1300
DLBCL	972	150

3. Results and Discussion

To validate the effectiveness of RS_SVM, we perform experiments on eight real gene expression datasets mentioned above. Three experiments are designed to validate the proposed method. In the first experiment, we computed testing error of RS_SVM and peer methods, including single SVM, KNN (K-nearest neighbor), CART (classification and regression tree), Bagging, and AdaBoost on eight datasets. Comparison of RS_SVM with the state-of-the-art methods in related literatures is also given. The second experiment explored influence of subspaces size by presenting the fluctuation of training error and testing error. In addition, sensitivity and specificity are also obtained at different subspace size. The last experiment shows the effectiveness of gene selection based on t-test.

The code is written in R-2.15.2, and all the packages are downloaded from the official site (https://www.r-project.org/). Table 3 gives a detailed description of the functions, the relative parameters, and packages used in experiments. Note that there is no training set and testing set partition on Colon Tumor, CNS, Ovarian, and DLBCL; we perform leave-one-out cross validation on these datasets.

3.1. Testing Error Comparison of RS_SVM and Other Methods. Table 4 shows testing error of RS_SVM and other peer methods on eight datasets. Testing error of each method is computed on the same dataset. To eschew the interference of randomness, values in Table 4 are the average of 50 iterations. It is clear that RS_SVM performs best on five datasets, that is, Breast Cancer, Lung Cancer, Prostate, Ovarian, and DLBCL. It also achieves good results on Leukemia dataset. Effect of aggregation is obvious by comparing RS_SVM with single SVM, since testing error of RS_SVM is lower on six datasets, and RS_SVM obtains the same result with single SVM on Colon Tumor. The only exception is CNS. For CNS, all the methods do not perform well, which probably was due to the special distribution of data.

Table 5 shows testing error of RS_SVM and the state-of-the-art methods in literatures. It is obvious that none of these methods is always the winner, since distribution or

Table 3: Function and package used in R.

Function	Package	Parameter
t.test()	stats	Confidence level of the interval is 0.95. Assume two variances are equal
svm()	e1071	Choose "radial" kernel; gamma is 1/dimension; epsilon is 0.1
knn()	class	Choose $k = 3$
rpart()	rpart	Choose method = "class"
ada()	ada	Use decision trees as base classifiers; iteration is 50; under exponential loss, type of boosting algorithm to perform is "discrete"
ipredbagg()	ipred	Use decision trees as base classifiers; number of bootstrap replications is 25

Table 4: Testing error comparison of RS_SVM and peer methods (%).

	RS_SVM	Single SVM	KNN	CART	AdaBoost	Bagging
Breast Cancer	**5.30**	15.79	47.37	31.58	10.53	31.58
Leukemia	5.89	26.47	**2.94**	8.82	41.18	8.82
Lung Cancer	**1.34**	9.40	2.68	9.40	51.01	9.40
Prostate	**0**	73.53	73.53	73.53	73.53	14.71
Colon Tumor	14.52	14.52	16.13	22.58	19.35	**11.29**
CNS	33.33	**31.67**	35.00	36.67	41.67	45.00
Ovarian	**1.19**	1.58	4.35	3.16	6.72	1.98
DLBCL	**4.26**	10.64	14.89	29.79	19.15	23.40

Table 5: Testing error comparison of RS_SVM and the state-of-the-art methods (%).

	Breast Cancer	Leukemia	Lung Cancer	Prostate	Colon Tumor	CNS	Ovarian	DLBCL
RS_SVM	**5.30**	5.89	1.34	**0**	14.52	33.33	1.19	4.26
Nanni et al. [28]	11.43	**0**	**0**	3.85	26.67	33.33	**0**	1.43
Ye et al. [29]	—	2.50	—	7.5	15.00	—	—	—
Liu et al. [30]	—	**0**	**0**	3.00	8.10	—	0.80	2
Tan and Gilbert [31]	—	8.90	6.80	26.50	4.90	11.7	—	—
Ding and Peng [32]	—	**0**	2.70	—	6.50	—	—	—
Bonilla Huerta et al. [33]	—	**0**	0.70	4.00	8.1	13.40	**0**	**0**
Cheng [34]	—	**0**	0.67	5.88	—	—	—	—
Paliwal and Sharma [35]	26.3	**0**	2.70	23.5	—	—	—	—
Bolón-Canedo et al. [10]	36.22	11.96	2.75	11.81	13.10	36.67	1.20	20.50
	46.56	4.11	**0**	41.87	16.19	30.00	0.8	6.50
	28.11	5.54	1.11	12.53	19.05	36.67	**0**	4.00
Porto-Díaz et al. [36]	21.05	**0**	0.67	20.59	10.00	25.00	**0**	**0**
Hu et al. [37]	—	—	12.50	19.30	9.70	—	—	—
	—	—	11.60	18.20	9.70	—	—	—
Nagi and Bhattacharyya [11]	26.51	7.55	18.12	47.06	5.60	**9.85**	1.11	
Pati and Das [38]	—	7.89	6.25	—	—	—	—	—
Dash et al. [39]	—	**0**	11.55	—	10.95	—	—	—
	—	0.45	**0**	—	**0**	—	—	—
	—	28.22	16	—	23.33	—	—	—
	—	0.41	0.95	—	0.31	—	—	—
Ghorai et al. [12]	18.79	5.48	3.62	9.84	17.23	—	—	—
Luo et al. [13]	—	2.07	—	—	18.60	—	—	6.00
	—	2.45	—	—	19.12	—	—	7.19

The state-of-the-art methods are indexed by the first author in literatures. "—" means that there are no corresponding results in the given literature.

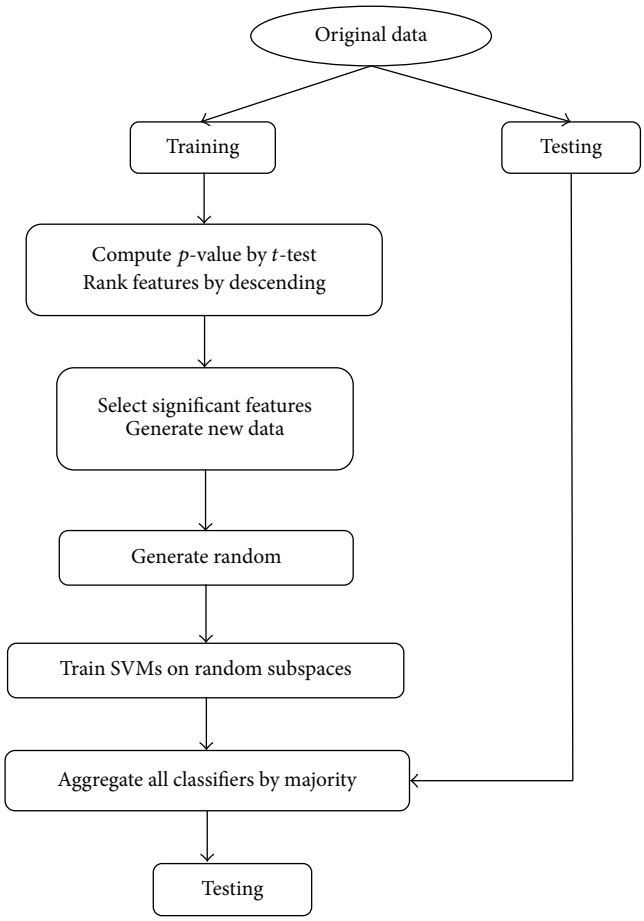


Figure 1: RS_SVM method.

correlation between gene features is diverse among different datasets. Each method has peculiar perspective for certain gene pattern. RS_SVM achieved the lowest testing error on Breast Cancer and Prostate and also relatively low testing error on the datasets of Leukemia, Lung Cancer, Ovarian, and DLBCL, which implies a good generalization performance.

In spite of good performances mentioned in Tables 4 and 5, an unsatisfied outcome is revealed on Colon Tumor and CNS. Possible reason might be traced to heterogeneity phenomenon appearing in the two datasets [37], which means greater variability existing in gene expression level between the categories. To visually describe the distribution, Figure 2 projects high-dimension data to two-dimension space by Principle Component Analysis (PCA). Heterogeneity phenomenon is obvious in Colon Tumor and CNS data. For CNS, distribution of "Class 1" is relatively concentrated and "Class 0" is more dispersing. Similar case happens on Colon Tumor. This suggests that RS_SVM is not suitable for heterogeneous data.

3.2. Influence of Subspace Size. Figure 3 shows training error and testing error with respect to subspace size. Breast Cancer, Leukemia, Lung Cancer, Ovarian, and DLBCL share nearly similar curve trend. Initially, both training error and testing error are high when subspace size is small, which indicates underfitting exists. With the increasing of subspace size, both errors converge to nearly zero and underfitting fades away. However, the convergence rate is different among different datasets. Ovarian data converges much slower than the other four datasets. Errors of Ovarian are not near zero until subspace size is almost 800.

For Colon Tumor, when training error is near zero, there is a small gap between training and testing errors. This indicates that slight overfitting exists. More severe overfitting exists on CNS, because there is an obviously large gap between training error and testing error when training error is converging to zero. The terrible overfitting may explain RS_SVM's high testing error in Tables 4 and 5.

For Prostate datasets, there is little variation on training error by increasing subspace size. Testing error, however, fluctuates dramatically, especially changing subspace size from 90 to 116. During this interval, testing error firstly drops down and minimum is obtained at the point when subspace size is set to 100, followed by rising up sharply, and finally tends to be steady. This phenomenon may be due to great differences between the distribution of training and testing set. As shown in Figure 4, tumor samples mainly concentrate in the left bottom in training set, while dispersing in the left in testing set. This indicates that the model generated on training set may not fit testing set well.

Figure 5 presents sensitivity and specificity with respect to subspace size. Sensitivity shows the ability to detect positives while specificity is the ability to reject negatives. To some extent, there is a trade-off between sensitivity and specificity. The best subspace size is a compromising value between sensitivity and specificity. For Breast Cancer, Leukemia, Lung Cancer, Ovarian, and DLBC, both sensitivity and specificity are high, which coincides with the low testing errors in Tables 4 and 5. Even though two curves of Colon Tumor are relatively steady, the whole level is not high. CNS dataset cannot achieve both high sensitivity and specificity, since when one rises up, the other drops down. The characteristic of Prostate dataset is also reflected in Figure 5. The sensitivity curve of Prostate rises up rapidly and then remains steady, but specificity curve drops down sharply when subspace size passes over the optimal value, which indicates that, with the increasing of subspace size, more and more tumor samples are predicted falsely.

3.3. Validation of Gene Selection by t-Test. The above experiments are performed on the datasets after gene selection via t-test, which is designed to reduce dimensionality and eliminate noise. In order to validate the effect of gene selection, we carry out experiment on datasets both with and without gene selection. Table 6 gives the testing error of RS_SVM on eight datasets. For the sake of contrast, parameters of two cases are all uniform. Size of subspace chooses the optimal value obtained in Table 2. It shows that gene selection improves classification performance obviously by reducing testing errors.

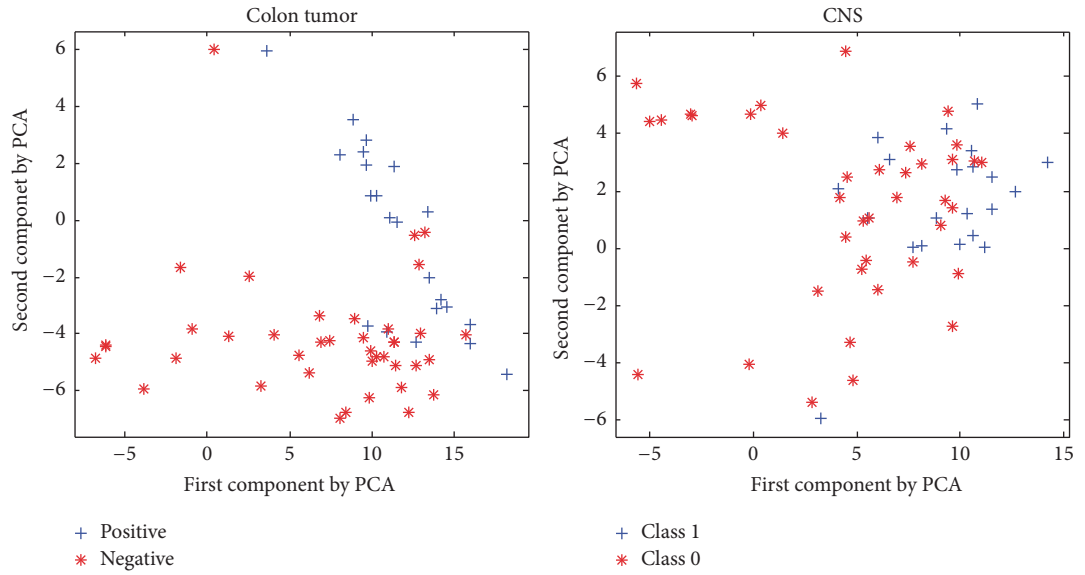


Figure 2: Scattering Colon Tumor and CNS data by Principle Component Analysis.

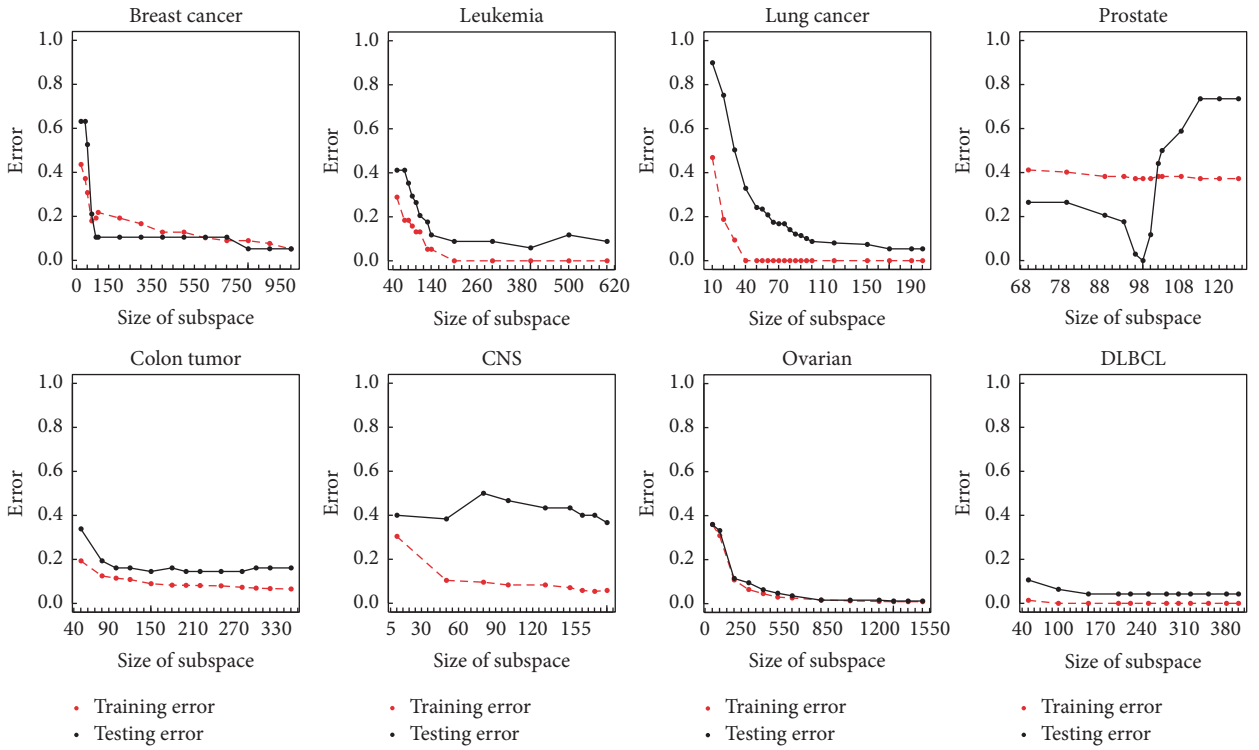


Figure 3: Variation of train error and test error with subspace size.

Table 6: Effect of gene selection based on t-test (%).

	Breast Cancer	Leukemia	Lung Cancer	Prostate	Colon Tumor	CNS	Ovarian	DLBCL
With selection	**5.30**	**5.89**	**1.34**	**0**	**14.52**	**33.33**	**1.19**	**4.26**
Without selection	63.16	41.18	3.36	26.47	35.48	35.00	3.20	44.68

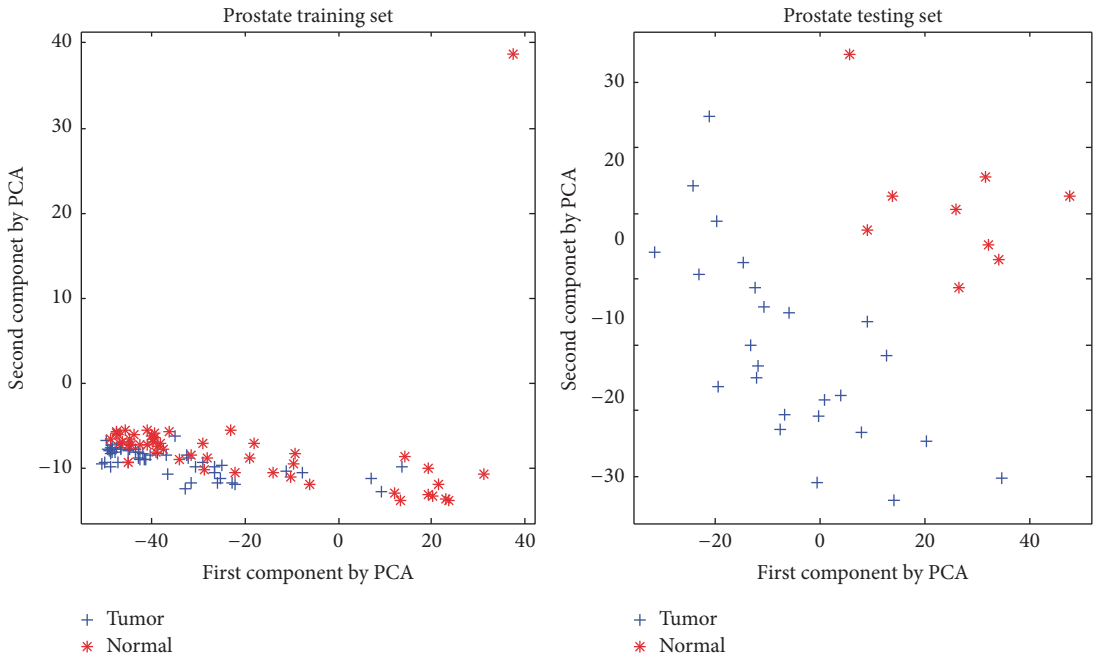


Figure 4: Scatter of training set and test set on Prostate based on the top two principle components.

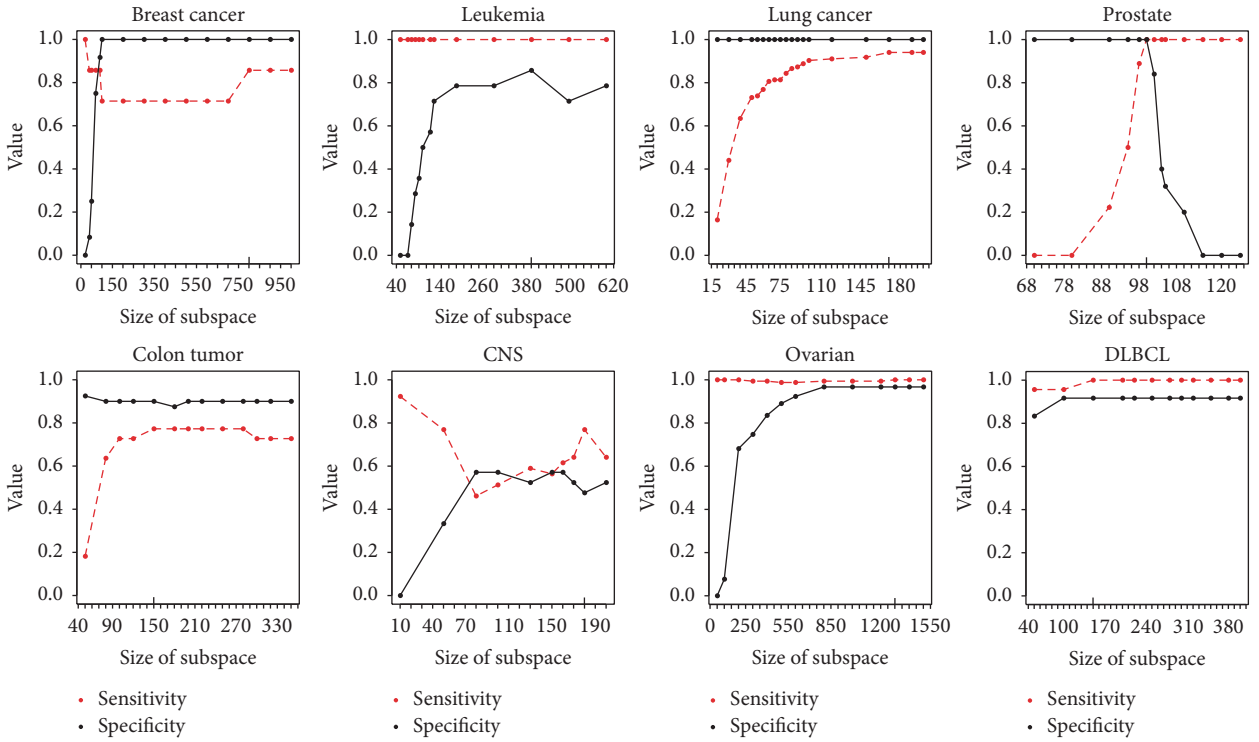


Figure 5: Variation of sensitivity and specificity with subspace size.

4. Conclusions

This work proposed a cancer classification method, termed RS_SVM, to analyze gene expression profiles. The robustness of SVM relies on the strong fundamentals of statistical learning theory and the technique can be extended to nonlinear discrimination by embedding the data in a nonlinear space using kernel functions. In pattern recognition systems, no single model exists for all pattern recognition problems and no single technique is applicable to all problems. Ensemble learning is to integrate several models for the same problem. Random subspace is one of the ensemble learning methods

and suitable for high-dimension data. For high-dimension gene expression data, only a small fraction of all genes is effective in performing certain diagnostic test. Hence, gene expression data analysis is confronted with enormous challenges for its characteristics, such as high dimensionality, small sample size, and low Signal-to-Noise Ratio. RS_SVM takes advantage of both subspace and SVM to handle the high-dimension and small sample problem in gene expression data, after obtaining the significant features through t-test, which could be regarded as prior knowledge to reduce the computing pressure. Experimental results on eight real gene expression profiles show that RS_SVM outperforms single SVM, KNN, CART, Bagging, AdaBoost, and 16 state-of-the-art methods in literatures. We also applied PCA on two gene expression profiles, where the experimental results are not satisfied, to probe the unsuitability. It suggests that RS_SVM is not suitable for heterogeneous data.

In RS_SVM, optimal values of subspace size and subspace number were obtained empirically, which was arduous and time-consuming. How to address this problem is still an open issue. We have collected next-generation sequencing gene expression data from TCGA and will continue this research on the new data.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Liying Yang conceived the project. Liying Yang, Zhimin Liu, Xiguo Yuan, and Junying Zhang designed the methodology. Liying Yang and Zhimin Liu performed the experiments, interpreted the results, and drafted the manuscript. Jianhua Wei, Xiguo Yuan, and Junying Zhang revised the manuscript.

Acknowledgments

This work was supported by the Natural Science Foundation of Shaanxi Province (CN) (2015JM6275), the Natural Science Foundation of China (61571341), and the Fundamental Research Funds for the Central Universities (JB160304).

References

[1] Q. M. Guo, "DNA microarray and cancer," *Current Opinion in Oncology*, vol. 15, no. 1, pp. 36–43, 2003.

[2] T. Zeng, R. Li, R. Mukkamala, J. Ye, and S. Ji, "Deep convolutional neural networks for annotating gene expression patterns in the mouse brain," *BMC Bioinformatics*, vol. 16, no. 1, article 147, 2015.

[3] V. Sachnev, S. Saraswathi, R. Niaz, A. Kloczkowski, and S. Suresh, "Multi-class BCGA-ELM based classifier that identifies biomarkers associated with hallmarks of cancer," *BMC Bioinformatics*, vol. 16, article 166, 2015.

[4] R. Li, W. Zhang, and S. Ji, "Automated identification of cell-type-specific genes in the mouse brain by image computing of expression patterns," *BMC Bioinformatics*, vol. 15, no. 1, article 209, 2014.

[5] M. Murtha, Z. Tokcaer-Keskin, Z. Tang et al., "FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells," *Nature Methods*, vol. 11, no. 5, pp. 559–565, 2014.

[6] C. L. Thompson, L. Ng, V. Menon et al., "A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain," *Neuron*, vol. 83, no. 2, pp. 309–323, 2014.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[8] X. Ren, Y. Wang, X.-S. Zhang, and Q. Jin, "IPcc: a novel feature extraction method for accurate disease class discovery and prediction," *Nucleic Acids Research*, vol. 41, no. 14, article e143, 2013.

[9] H. Zhang, H. Wang, Z. Dai, M.-S. Chen, and Z. Yuan, "Improving accuracy for cancer classification with a new algorithm for genes selection," *BMC Bioinformatics*, vol. 13, no. 1, article 298, 20 pages, 2012.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recognition*, vol. 45, no. 1, pp. 531–539, 2012.

[11] S. Nagi and D. K. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 3, pp. 159–173, 2013.

[12] S. Ghorai, A. Mukherjee, S. Sengupta, and P. K. Dutta, "Cancer classification from gene expression data by NPPC ensemble," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 659–671, 2011.

[13] L. Luo, L. Ye, M. Luo, D. Huang, H. Peng, and F. Yang, "Methods of forward feature selection based on the aggregation of classifiers generated by single attribute," *Computers in Biology and Medicine*, vol. 41, no. 7, pp. 435–441, 2011.

[14] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[15] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[16] G. Daxa, K. Ankit, and G. Monali, "Classification of gene expression data by gene combination using fuzzy logic," *International Journal of Advance Research and Innovative Ideas in Education*, vol. 1, no. 2, pp. 43–48, 2015.

[17] Y. Kim, Y. Yoon, F. Liu, D. Lee, R. Lagoa, and S. Kumar, "A genetic filter for cancer classification on gene expression data," *Bio-Medical Materials and Engineering*, vol. 26, supplement 1, pp. S1993–S2002, 2015.

[18] M. Vosooghifard and H. Ebrahimpour, "Applying Grey Wolf Optimizer-based decision tree classifer for cancer classification on gene expression data," in *Proceedings of the 5th International Conference on Computer and Knowledge Engineering (ICCKE '15)*, pp. 147–151, IEEE, Mashhad, Iran, October 2015.

[19] K. Buza, "Classification of gene expression data: a hubness-aware semi-supervised approach," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 105–113, 2016.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[21] A. Bertoni, R. Folgieri, and G. Valentini, "Bio-molecular cancer prediction with random subspace ensembles of support vector machines," *Neurocomputing*, vol. 63, pp. 535–539, 2005.

[22] L. I. Kuncheva, J. J. Rodríguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 531–542, 2010.

[23] X. Li and H. Zhao, "Weighted random subspace method for high dimensional data classification," *Statistics and its Interface*, vol. 2, no. 2, pp. 153–159, 2009.

[24] J. Li and H. Liu, http://datam.i2r.a-star.edu.sg/datasets/krbd/.

[25] L. J. Van't Veer, H. Dai, M. J. Van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[26] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[27] G. J. Gordon, R. V. Jensen, L.-L. Hsiao et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[28] L. Nanni, S. Brahnam, and A. Lumini, "Combining multiple approaches for gene microarray classification," *Bioinformatics*, vol. 28, no. 8, pp. 1151–1157, 2012.

[29] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using uncorrelated discriminant analysis for tissue classification with gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 4, pp. 181–190, 2004.

[30] B. Liu, Q. Cui, T. Jiang, and S. Ma, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC Bioinformatics*, vol. 5, article 136, 2004.

[31] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 2, no. 3, pp. S75–S83, 2003.

[32] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[33] E. Bonilla Huerta, B. Duval, and J.-K. Hao, "A hybrid LDA and genetic algorithm for gene selection and classification of microarray data," *Neurocomputing*, vol. 73, no. 13–15, pp. 2375–2383, 2010.

[34] Q. Cheng, "A Sparse learning machine for high-dimensional data with application to microarray gene analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 4, pp. 636–646, 2010.

[35] K. K. Paliwal and A. Sharma, "Improved direct LDA and its application to DNA microarray gene expression data," *Pattern Recognition Letters*, vol. 31, no. 16, pp. 2489–2492, 2010.

[36] I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, and O. Fontenla-Romero, "A study of performance on microarray data sets for a classifier based on information theoretic learning," *Neural Networks*, vol. 24, no. 8, pp. 888–896, 2011.

[37] P. Hu, S. B. Bull, and H. Jiang, "Gene network modular-based classification of microarray samples," *BMC bioinformatics*, vol. 13, supplement 10, p. S17, 2012.

[38] S. K. Pati and A. K. Das, "Gene selection and classification rule generation for microarray dataset," in *Advances in Computing and Information Technology*, vol. 178 of *Advances in Intelligent Systems and Computing*, pp. 73–83, Springer, Berlin, Germany, 2013.

[39] S. Dash, B. Patra, and B. Tripathy, "A hybrid data mining technique for improving the classification accuracy of microarray data set," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 2, pp. 43–50, 2012.

[40] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[41] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[42] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[43] E. F. Petricoin, A. M. Ardekani, B. A. Hitt et al., "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, pp. 572–577, 2002.

[44] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[45] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.